THE AMERICAN MATHEMATICAL
MONTHLY

Volume 106, Number 7 August–September 1999

AN OFFICIAL PUBLICATION OF THE MATHEMATICAL ASSOCIATION OF AMERICA

NOTICE TO AUTHORS

The MONTHLY publishes articles, as well as notes and other features, about mathematics and the profession. Its readers span a broad spectrum of mathematical interests, and include professional mathematicians as well as students of mathematics at all collegiate levels. Authors are invited to submit articles and notes that bring interesting mathematical ideas to a wide audience of MONTHLY readers.

The MONTHLY's readers expect a high standard of exposition; they expect articles to inform, stimulate, challenge, enlighten, and even entertain. MONTHLY articles are meant to be read, enjoyed, and discussed, rather than just archived. Articles may be expositions of old or new results, historical or biographical essays, speculations or definitive treatments, broad developments, or explorations of a single application. Novelty and generality are far less important than clarity of exposition and broad appeal. Appropriate figures, diagrams, and photographs are encouraged.

Notes are short, sharply focussed, and possibly informal. They are often gems that provide a new proof of an old theorem, a novel presentation of a familiar theme, or a lively discussion of a single issue.

Articles and Notes should be sent to the Editor:

ROGER A. HORN
1515 Mineral Square, Room 142
University of Utah
Salt Lake City, UT 84112

Please send your email address and 3 copies of the complete manuscript (including all figures with captions and lettering), typewritten on only one side of the paper. In addition, send one original copy of all figures without lettering, drawn carefully in black ink on separate sheets of paper. Authors who use LaTeX are urged to use article.sty and its standard environments with no custom formatting

Letters to the Editor on any topic are invited; please send to the MONTHLY's Utah office. Comments, criticisms, and suggestions for making the MONTHLY more lively, entertaining, and informative are welcome.

See the MONTHLY section of MAA Online for current information such as contents of issues and descriptive summaries of forthcoming articles:

http://www.maa.org/

Proposed problems or solutions should be sent to:

DANIEL ULLMAN, MONTHLY Problems
Department of Mathematics
The George Washington University
2201 G Street, NW, Room 428A
Washington, DC 20052

Please send 2 copies of all problems/solutions material, typewritten on only one side of the paper.

Reprint permission:
DONALD ALBERS, Director of Publications

Advertising Correspondence:
Dave Riska, *driska@maa.org*

Change of address, missing issues inquiries, and other subscription correspondence:
MAA Service Center, *maahq@maa.org*

All at the address:

The Mathematical Association of America
1529 Eighteenth Street, N.W.
Washington, DC 20036

Recent copies of the MONTHLY are available for purchase through the MAA Service Center,
maahq@maa.org, 1-800-331-1622

Microfilm Editions: University Microfilms International, Serial Bid coordinator, 300 North Zeeb Road, Ann Arbor, MI 48106.

The AMERICAN MATHEMATICAL MONTHLY (ISSN 0002-9890) is published monthly except bimonthly June-July and August-September by the Mathematical Association of America at 1529 Eighteenth Street, N.W., Washington, DC 20036 and Montpelier, VT.

The Hopping Hoop Revisited

Timothy Pritchett

In this MONTHLY, T. Tokieda [1] recently provided a novel discussion of the hopping hoop problem described by Littlewood [2]. Unfortunately, the analysis in [1] is incorrect. We show that if the "no slipping" condition imposed in [1] and [2] is strictly adhered to, the hoop never becomes airborne. Essential for the observed hop of the hoop is a phase in which there is slippage at the point of contact of the hoop with the supporting surface. In order to expose this and other subtleties not considered in [1], we treat a somewhat more general problem. The view of Littlewood's hopping hoop problem presented here is consciously different from that given in [1]. In offering this alternative way of thinking about Littlewood's hopping hoop, we hope to deepen the reader's intuitive understanding of precisely why the hoop hops.

We consider here a rigid circular hoop of radius R and mass $(1 - \lambda)M$ that, at least initially, rolls without slipping along a flat surface. An additional object of mass λM is rigidly attached to the rim of the hoop, with the point of attachment coinciding with the centroid of the object. Here, M is the total mass of the combined system consisting of the hoop and the attached object. The parameter λ is the ratio of the object mass to the total; in the limit $\lambda \to 1$, the problem reduces to the case of the massless hoop considered by Tokieda and Littlewood. The center of mass of the combined system, shown as a dot in Figure 1, lies a distance

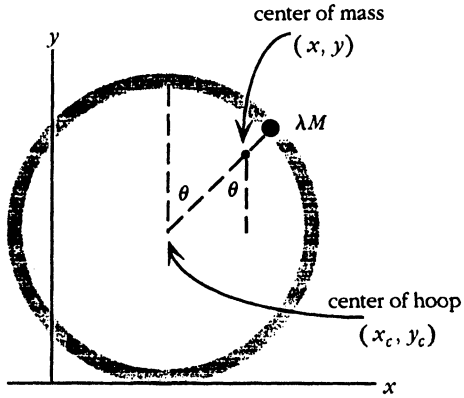


Figure 1. The hopping hoop: a circular hoop of mass $(1 - \lambda)M$ with an attached object of mass λM.

λR from the center of the hoop and is situated on the line segment connecting the center of the hoop with the attached object. Let $\theta(t)$ be the angle which this line segment makes with the vertical and let $x(t)$ and $y(t)$ be, respectively, the horizontal and vertical coordinates of the center of mass of the system.

The three coordinates $x(t)$, $y(t)$, and $\theta(t)$ are not independent, but are subject to various constraints:

1. Since the hoop is assumed to be a rigid object that suffers no deformation as it rolls, it is always true that $(x(t) - x_c(t))^2 + (y(t) - y_c(t))^2 = \lambda^2 R^2$, or

equivalently,

$$x(t) - x_c(t) = R\lambda \sin \theta(t) \tag{1}$$

$$y(t) - y_c(t) = R\lambda \cos \theta(t) \tag{2}$$

where $x_c(t)$ and $y_c(t)$ denote, respectively, the horizontal and vertical coordinates of the center of the hoop.

2. Since the supporting surface is also assumed to be rigid, the vertical height of the center of the hoop must satisfy $y_c(t) = R$ so long as the hoop and the surface are in contact.
3. As long as there is no slipping, the horizontal position of the center of the hoop and the angular coordinate $\theta(t)$ are related by $x_c(t) = R\theta(t)$, where we have tacitly set $x_c(0) = R\theta(0) = 0$.

Taken together, these three constraints imply that as long as the hoop rolls without slipping along the supporting surface, the center of mass of the system moves along a curtate (shortened) cycloid, also known as a trochoid.

Every constraint is maintained by a corresponding force. In particular, constraints 2 and 3 result from the force of contact between the hoop and the supporting surface: the condition $y_c(t) = R$ is maintained by the so-called normal force, i.e., the component of the contact force acting perpendicular to the surface, while the "no slipping" constraint, $x_c(t) = R\theta(t)$, is maintained by friction, i.e., by the component of the contact force acting parallel to the surface. Constraint 1, the "rigid hoop" constraint, is maintained by forces within the hoop itself; because these are internal to the system under consideration, they need not concern us here.

All forces external to the system, including those responsible for maintaining the constraints just enumerated, appear in the equations of motion of the center of mass of the system, which are obtained from Newton's Second Law:

$$\ddot{x}(t) = f(t) \tag{3}$$

$$\ddot{y}(t) = n(t) - g \tag{4}$$

Here, g, $n(t)$, and $f(t)$ represent, respectively, the force per unit mass (acceleration) due to gravity, to the normal component of the contact force, and to friction. In order to simplify the equations, we divide all forces by M, the total mass of the system. For brevity we continue to refer to the resulting quantities as "forces" even though they are actually accelerations, i.e., forces per unit mass. When the hoop loses contact with the supporting surface, $n(t) = f(t) = 0$, and the equations (3) and (4) describe a parabolic trajectory.

The most general motion of a rigid body consists of a translation of the center of mass, combined with a rotation about an axis containing the center of mass. In the present case, this rotation is described by the angular variable $\theta(t)$, whose time evolution is governed by the following equation, obtained by considering the torques about the axis perpendicular to the plane of Figure 1 and passing through the center of mass of the system:

$$\frac{I}{M}\ddot{\theta}(t) = n(t)[x(t) - x_c(t)] - f(t)y(t) \tag{5}$$

In this expression, I is the moment of inertia of the system (hoop + object) about the axis through the system center of mass. It is given by $I = MR^2\{1 - \lambda^2 + \lambda\epsilon\}$, where the last term in brackets requires some comment. Unless the attached object is a point mass (a mathematical idealization), it has a non-vanishing moment

of inertia about any axis through its centroid. That "internal" moment of inertia is proportional to the mass λM of the attached object and to the square of some length parameter related to the dimensions of the object in the plane perpendicular to the axis of rotation; the moment of inertia is of the form $\lambda M b^2$. For example, for a cylinder of uniform mass density (a battery!), attached to the hoop so that its longitudinal axis is aligned with the hoop, $b = L/\sqrt{12}$, where L is the length of the cylinder; for the same cylinder, attached to the hoop in such a way that its longitudinal axis is perpendicular to the plane of the hoop, $b = r/\sqrt{12}$, where r is the cylinder radius; and so on. For a point mass, $b = 0$. Instead of b, however, we choose to work with the dimensionless parameter $\epsilon = (b/R)^2$. Since $\epsilon \ll 1$ in almost all cases, one might wonder why we make a large fuss over a quantity that, in practice, is negligible small. The answer is that we wish to examine the limit $\lambda \to 1$, in which the mass of the hoop is negligible relative to that of the attached object. In this limit, the center of mass of the system coincides with the centroid of the object, and the system's angular momentum relative to its center of mass resides entirely in the attached object spinning about its centroid; this angular momentum is zero if the corresponding moment of inertia vanishes, as is the case if we set to zero both the mass of the hoop and the "internal" rotational inertia of the attached object. Alternately, we observe that (5) makes sense only so long as $I \neq 0$, and, in the case of a massless hoop, I is nonvanishing only if we account for the rotational inertia of the attached object about its own centroid. The importance of (5) becomes clearer when we relax the constraint that the hoop rolls entirely without slipping.

Returning to (5), we make the following important observation: When hoop and supporting surface part company, the contact forces $n(t)$ and $f(t)$ vanish and (5) reduces to $\ddot{\theta}(t) = 0$. Thus, the hoop simply rotates about its center of mass at a constant angular speed equal to its angular speed at the instant the contact forces went to zero. At the same time, equations (3) and (4), which describe the translational motion of the center of mass, reduce to the equations for a body falling freely with a constant downward acceleration. This simultaneous free fall/free rotation continues until the height $y_c(t)$ of the center of the hoop decreases once again to R, at which time $n(t)$ acquires the positive value required to enforce constraint 2. In this way, we arrive at an alternative view of why the hoop hops: it is simply rotating about its center of mass, as that center of mass falls freely under the influence of gravity. Littlewood's hopping hoop is reminiscent of a good high jumper, who is skilled at rotating and "deforming" her body in such a way that all of its component parts clear the bar, even as her center of mass passes *under* it!

The motion of the system is completely determined by equations (3), (4), and (5), along with the relevant constraints and initial conditions. We have already noted that constraints 1, 2, and 3 are simultaneously satisfied only if the center of mass of the system moves along the trochoid:

$$x(t) = R(\theta(t) + \lambda \sin \theta(t)) \tag{6}$$

$$y(t) = R(1 + \lambda \cos \theta(t)) \tag{7}$$

By combining equations (3) through (7) one may obtain a single second-order differential equation for one of the variables ($\theta(t)$, say) from which the motion of the system may be determined. The result will be valid as long as the constraints embodied by (6) and (7) hold, i.e., as long as the hoop rolls without slipping along the supporting surface. Alternately, one may observe that the constraints imply

that $x(t) = y(t)\dot{\theta}(t)$ and $\dot{y}(t) = (x(t) - x_c(t))\dot{\theta}(t)$, and use these relations to combine (3), (4), and (5) into a single equation, from which the forces of constraint $n(t)$ and $f(t)$ are absent and which, more importantly, may be written as a total derivative. In this way, one arrives at the equation of energy conservation. Physically, the absence of the forces of constraint from the equation of energy conservation reflects the fact that forces of constraint do no net work because they act perpendicular to the configuration space of the system. If now we assume, as Tokeida does, that the attached mass initially moves horizontally with speed v_0, energy conservation requires

$$\tfrac{1}{2}M\left(\dot{x}^2 + \dot{y}^2\right) + \tfrac{1}{2}I\dot{\theta}^2 + Mgy = M\left((1+\lambda)\left(\tfrac{1}{4}v_0^2 + gR\right) + \tfrac{1}{8}v_0^2\lambda\epsilon\right),$$

where the additional terms not seen in the corresponding equation in [**1**] arise from the rotational kinetic energy of the (massive) hoop and attached (extended) object. Adding the trochoid constraint, equations (6) and (7), we obtain

$$\dot{\theta}(t)^2 = \frac{g}{R}\frac{\left((1+\lambda+\lambda\epsilon/2)c + 2\lambda\sin^2\left(\theta(t)/2\right)\right)}{1+\lambda\cos\left(\theta(t)\right) + \lambda\epsilon/2} \tag{8}$$

where $c = v_0^2/(4gR)$. This result may be used in conjunction with (4) to compute $\ddot{y}(\theta(t))$, as in [**1**]. Tokieda observes that the hop must occur at the first value of $\theta(t)$ for which the parabolic trajectory of the center of mass in free fall departs from above from the trochoidal trajectory imposed by the rolling hoop, i.e., the hop occurs when $\ddot{y}(\theta(t)) = -g$. But, referring to (4), this is precisely when $n(t) = 0$, i.e., when the normal component of the force of contact between the hoop and the supporting surface vanishes, a condition that certainly *seems* to suggest that the hoop has just lost contact with the supporting surface and is about to become airborne. That is not the case, as we now prove.

Combining (4), (7), and (8), we obtain the following rather formidable expression for the normal force n as a function of $\xi = \cos\theta(t)$.

$$n(\xi) = \frac{g}{(\epsilon\lambda + 2\xi\lambda + 2)^2}p(\xi), \qquad \text{where}$$

$$\begin{aligned} p(\xi) = {} & (1 - c\lambda\xi)\lambda^2\epsilon^2 \\ & - \epsilon\lambda\left(\left(c\xi^2 + 2c\xi + 2\xi + c - 3\xi^2 + 1\right)\lambda^2 + 4(c-1)\xi\lambda - 4\right) \\ & - 2\left(\left(c\xi^2 + \xi^2 + c - 2\xi^3 + 1\right)\lambda^3\right. \\ & \left. + \left(c\xi^2 + 2c\xi + 2\xi + c - 5\xi^2 + 1\right)\lambda^2 + 2(c-2)\xi\lambda - 2\right) \end{aligned} \tag{9}$$

We now specialize to the case of a massless hoop ($\lambda = 1$) and determine the angle θ_1 for which $n \to 0$ by finding the roots of the cubic polynomial $p(\xi)$ in (9). The result is most illuminating if we express it as a series expansion in the "inertia parameter" ϵ. The physically relevant root is

$$\cos\theta_1 = c - \frac{(1-c)}{(c+1)}\left(\frac{(c+3)}{4}\epsilon + \frac{1}{(c+1)^2}\epsilon^2\right) + O(\epsilon)^3. \tag{10}$$

If the attached object is a point mass ($\epsilon = 0$), we recover from (10) the result given in [**1**] for the critical angle θ_1 at which, it is claimed, the hoop loses contact with the supporting surface. If that is indeed the case, then the subsequent motion of the system must consist of a translation of the center of mass along the parabolic trajectory corresponding to free fall,

$$x(t) = x_1 + \dot{x}(t - t_1) \tag{11}$$

$$y(t) = y_1 + \dot{y}_1(t - t_1) - \tfrac{1}{2}g(t - t_1)^2, \tag{12}$$

combined with a rotation about the center of mass at constant angular speed $\dot{\theta}_1$:

$$\theta(t) = \theta_1 + \dot{\theta}_1(t - t_1). \tag{13}$$

Here, (x_1, y_1) and $(\dot{x}_1, \dot{y}_1)$ are the Cartesian coordinates of the position and velocity of the center of mass at the instant t_1 at which the normal force vanishes, and $\dot{\theta}_1$ is the corresponding angular speed of rotation. Series expansions of these quantities are obtained from (10), in conjunction with (6), (7), and (8). If one now substitutes (12) and (13) into (2), replacing y_1, $\dot{y}_1$, θ_1, and $\dot{\theta}_1$ by their expansions in ϵ, one obtains the following expression for the height $y_c[t]$ of the center of the hoop at times $t > t_1$:

$$\frac{y_c(\tau)}{R} = 1 - \frac{(1-c)\big((\epsilon+4)c^2 + 8c + 3\epsilon + 4\big)}{8(c+1)^2}\tau^2 + O(\epsilon)^2 + O(\tau)^3, \tag{14}$$

where $\tau = \sqrt{g/R}(t - t_1)$. Now, c must be strictly less than unity, or else the requirement that the hoop initially rolls without slipping cannot be satisfied; for $c \geq 1$ (i.e., for $v_0 \geq \sqrt{4gR}$), the hoop "glides" immediately without rolling, as Tokieda points out [**1**]. Since $c < 1$, the term of order τ^2 in (14) is negative. This means that an instant after the normal force between the hoop and the supporting surface goes to zero, the center of the hoop is moving not upward, but downward, a result that one could have obtained equally well by computing the acceleration $\ddot{y}_c(t_1)$! Thus, even though at t_1 the normal component of the contact force is instantaneously zero, it cannot remain zero for a nonvanishing duration. This is because at t_1 the center of the hoop is accelerating downward, and a positive normal force immediately results to enforce constraint 2 and prevent the height of the center of the hoop from falling below R. Equation (14) shows that a massless hoop cannot get off the ground if we insist that there be no slipping. In particular, this is true for the special case considered in [**1**] ($\epsilon = 0$: attached object is a point mass) as well as for Littlewood's original, formulation ($\epsilon = c = 0$: attached point mass with zero initial velocity) [**2**]. In the case of a massive hoop ($\lambda < 1$), the problem must be treated numerically, but the result is the same: as long as we require that the hoop roll entirely without slipping, it will never become airborne.

Even if the forces of constraint vanish instantaneously, the hoop and the supporting surface can part company only if $\dot{y} + R\dot{\theta}\sin\theta > 0$, i.e., only if the speed $R\dot{\theta}\sin\theta$ at which the center of the hoop rises (due to the rotation of the system about its center of mass) *exceeds* the speed $|\dot{y}|$ at which the center of mass falls under the influence of gravity. However, as long as we stipulate that the hoop rolls entirely without slipping, imposing by fiat the trochoid constraint, then it follows from (7) that the magnitude of the quantities $R\dot{\theta}\sin\theta$ and $\dot{y}$ are *equal*. Right on the money is Littlewood's [**2**] comment that, in practice, the hoop slips first (before hopping).

At what point does the hoop begin to slip? By combining equations (3), (6), and (8), one obtains the following expression for the force f responsible for maintaining the "no slipping" condition, constraint 3.

$$f(\xi) = -\frac{g\lambda\sqrt{1-\xi^2}}{(\lambda\epsilon + 2\lambda\xi + 2)^2}q(\xi), \qquad \text{where}$$

$$q(\xi) = 2\epsilon\lambda^2 + (2\lambda - 3\epsilon\lambda + c(\epsilon\lambda + 2\lambda + 2) - 6)\xi\lambda + 2\lambda - \epsilon\lambda \tag{15}$$
$$+ c(\epsilon^2\lambda^2 + 2\epsilon\lambda^2 + 3\epsilon\lambda + 2\lambda + 2) - 4\lambda^2\xi^2 - 2$$

We remind the reader that (9) and (15) represent, respectively, the vertical and horizontal components of the contact force necessary at any given angle $\theta = \arccos(\xi)$ to keep the center of mass moving on the trochoidal trajectory dictated by constraints 1, 2, and 3. It is natural to ask whether the required force is indeed physically available.

The normal force n is limited only by the resistance to tensile stress of the hoop and the supporting surface; in assuming that both the hoop and the surface are infinitely rigid, not subject to deformations of any kind, we allow n to be as large as is necessary to maintain constraint 2. In the real world, of course, no object is infinitely rigid. In the limit of small strains, however, the resulting deformations are elastic—energy-conserving—and the potential energy stored as elastic strain in the hoop and/or supporting surface just prior to the loss of contact provides an additional source of propulsion for the hop.

The frictional force f, on the other hand, has a maximum value, and if this is exceeded, slipping occurs. It is known experimentally that the maximum possible magnitude of the force of static friction existing at the point of contact of two surfaces is proportional to the magnitude of the normal force of contact exerted by one surface on the other:

$$|f| \leq |f_{\mathrm{MAX}}| = \mu_s |n| \tag{16}$$

The constant of proportionality, μ_s, is known as the *coefficient of static friction* and it depends on the nature of the surfaces that are in contact. For most materials, μ_s ranges from around 0.01 to 1.5 **[3]**. Once the point of contact between the hoop and the supporting surface begins to slip, the magnitude of the friction force is proportional to the magnitude of the normal force: $|f(t)| = \mu_k |n(t)|$. The relevant constant of proportionality, the coefficient of kinetic friction μ_k, is always less than the coefficient of static friction μ_s, for any pair of surfaces in contact. This is because the "cold-welding" that serves to bind the surfaces together in the static case cannot occur when there is relative motion between the two surfaces **[3]**. Once slipping begins, the friction force is directed so as to oppose the relative motion of the two surfaces involved. For the skidding hoop,

$$f(t) = \mu_k n(t) \operatorname{sgn}\left(R\dot{\theta}(t) - \dot{x}_c(t)\right), \tag{17}$$

where $\operatorname{sgn}(x)$ is the sign function and $n(t) \geq 0$. Specifically, $f(t)$ is positive (directed to the right) if the hoop is skidding forward, i.e., if the motion of the hoop relative to the point of contact is directed to the left, and $f(t)$ is negative (directed to the left) if the hoop is skidding backward, i.e., if the motion of the hoop relative to the point of contact is directed to the right.

The hoop begins to slip at the minimum angle for which equality holds in (16), i.e., for the minimum (real) root of $\{f(\cos\theta)\}^2 - \{\mu_s n(\cos\theta)\}^2 = 0$, where the forces of constraint n and f are given by (9) and (15), respectively. For example, in the case of a massless hoop ($\lambda = 1$) with attached point mass ($\epsilon = 0$), θ_{slip} is the lesser of $\arccos(c)$ and $2\arctan(\mu_s)$. The subsequent motion of the system is computed using (17) and numerically solving the coupled differential equations (3) and (5) subject to constraints 1 and 2. The former constraint (no deformations) translates simply into the twin conditions expressed by equations (1) and (2). Considerably more tedious to implement is the latter constraint, which must be enforced explicitly at each time step in the integration. One proceeds as follows: A provisional value of the normal force $n(t)$ is obtained from $n(t - \Delta t)$ by extrapola-

tion. From $n(t)$, $x(t)$, $y(t)$, $\theta(t)$, $\dot{x}(t)$, $\dot{y}(t)$, and $\dot{\theta}(t)$ one computes a provisional value of $y_c(t + \Delta t)$. If $y_c(t + \Delta t) < R$, one adjusts $n(t)$ to give $y_c(t + \Delta t) = R$. The revised n is then used to compute the values of x, y, θ, $\dot{x}$, $\dot{y}$, and $\dot{\theta}$ at time $t + \Delta t$. The procedure continues until $y_c(t)$ is strictly greater than R, at which point the hoop is airborne and the subsequent motion of the system is given by (11)–(13), where the subscript "1" now refers to values at the final step in the numerical procedure, i.e., at the instant contact between hoop and supporting surface is lost.

The implementation of the procedure just described is complicated by the fact that comparisons of approximate (floating point) quantities can be performed only to a certain tolerance. Thus, the stopping condition that y_c be strictly greater than R is, in practice, $y_c - R > \delta$; that is, we consider the hoop to be in contact with the supporting surface and, more importantly, we continue to impose constraints 1 and 2, until the point of contact clears the surface by at least δ. Since the parameter δ effectively *defines* when the hoop is considered to be airborne, the amount of time during which the hoop rolls while slipping prior to takeoff depends on δ, and so too does the height of the hop. Roughly speaking, this is because the longer the "rolling with slipping" phase prior to takeoff, the greater the takeoff speed that the hoop can build up—more precisely: the greater the amount by which $|R\dot{\theta} \sin \theta|$ can exceed $|\dot{y}|$. Here, $|\dot{y}|$ is the speed at which the center of mass falls under the influence of gravity, while $|R\dot{\theta} \sin \theta|$ is the speed at which the center of the hoop rises as a result of the rotation of the system about its centroid. Numerical simulations confirm that the height of the hop decreases with δ, so one might expect in the limit $\delta \to 0$ to observe no hop at all. However, the $\delta \to 0$ limit is not only in conflict with physical reality, it violates even the assumptions of this admittedly idealized problem! The point is simply this: Even if we were able to perform our numerical computations to infinite precision, δ could still not be made arbitrarily small. It is, after all, microscopic imperfections in the hoop and the supporting surface that gave rise to friction in the first place, and the scale of these imperfections provides a physical lower bound to δ. In the present case, we began with the assumption that hoop and surface are sufficiently "rough" that the former would initially roll without slipping over the latter. This precludes $\delta \to 0$.

The preceding discussion might lead one to doubt that the hop can actually be observed in practice. This is not the case: the hop is real, as the photograph in Figure 2 shows.

Figure 3 depicts schematically the results of a typical numerical simulation using parameter values for a system that one could potentially realize experimentally: the hoop is *not* massless, and the attached object is *not* an idealized point mass. Shown are the trajectory of the center of mass (black line) and the position of the center of mass at the onset of slipping and at the instant the hoop loses contact with the supporting surface (open circles). Dots indicate the positions of the attached object (large dots) and the center of the hoop (small dots) at twenty equally spaced times. Spokes (light gray lines) connecting simultaneous positions of the object and the center of the hoop have been added as an aid in visualization. The figure was generated using $\lambda = 0.95$, $\epsilon = 0.01$, and $c = 0.1$, the latter value corresponding to the attached object moving horizontally with an initial speed of 2 meters/sec. The coefficient of static and kinetic friction between the hoop and supporting surface were taken to have values 1.0 and 0.8, respectively. The simulation generating the figure used $\delta = R/100$; in simulations employing smaller values of δ, the hop is less evident, but it is always present.

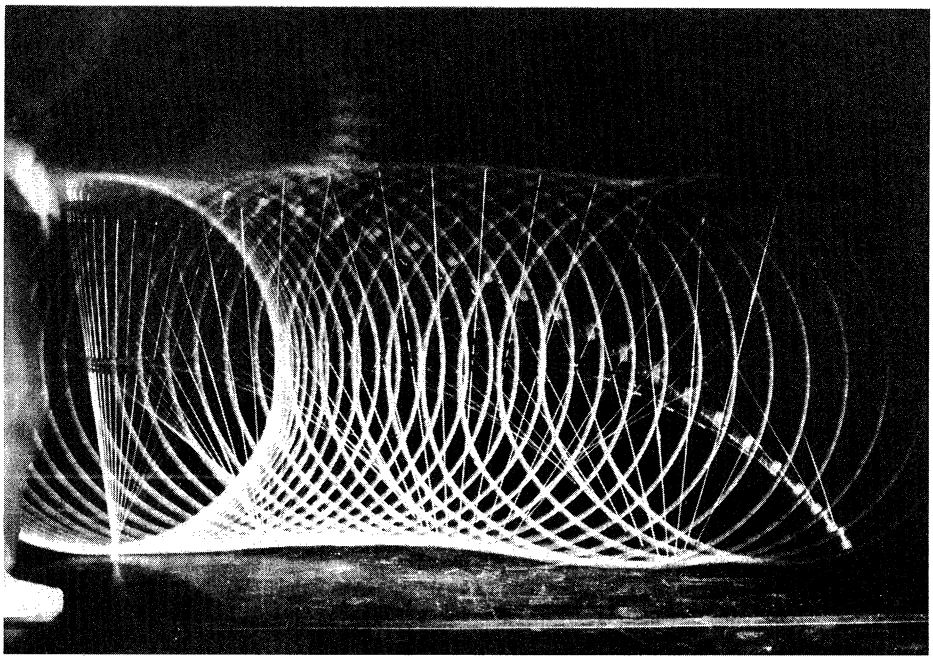


Figure 2. A stroboscopic photo by Dan Schwalbe and Stan Wagon shows a small hop of the hoop, which is a plastic hula hoop and four brass rods.

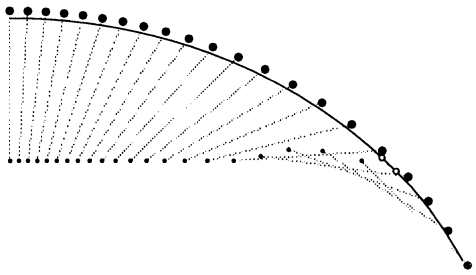


Figure 3. Simulated hop of a real hoop: results of a numerical simulation using $\lambda = 0.95$, $\epsilon = 0.01$, and $c = 0.1$ and coefficients of friction $\mu_s = 1.0$ and $\mu_k = 0.8$. The large dots indicate the positions of the attached object and the small dots indicate the corresponding positions of the center of the hoop. The trajectory of the center of mass of the system is indicated by the black line; the two open circles along the center of mass trajectory indicate the position of the center of mass at the onset of slipping and at the instant of loss of contact with the supporting surface.

We note in passing that if the hoop is massless ($\lambda = 0$) and the attached object is a point mass ($\epsilon = 0$), (8) can be integrated in closed form to give the following rather cute solution for $\theta(t)$:

$$\theta(t) = 2\arcsin\left[\sqrt{c}\,\sinh\left[\frac{t}{2}\sqrt{\frac{g}{R}}\right]\right] = 2\arcsin\left[\frac{v_0}{2\sqrt{gR}}\sinh\left[\frac{t}{2}\sqrt{\frac{g}{R}}\right]\right].$$

In his original essay [2], Littlewood gives his discussion of the hopping hoop a somewhat philosophical turn by posing a question to which we, in closing, now turn. Wonders Littlewood: Is the behavior of the hoop intuitive? We have demonstrated that there can be no hop without some slipping of the hoop. This

conclusion may at first appear somewhat surprising. We hope, however, that the reader now agrees that the phenomenon of hopping can be quite intuitive, at least if it is explained in general terms. In this instance, as in countless others, the devil is in the details.

ACKNOWLEDGMENT. Thanks for this article are due to Stan Wagon, who encouraged the author to develop a computer animation of the hopping hoop, and who supplied the stroboscopic photograph in Figure 2.

REFERENCES

1. Tadashi F. Tokieda, The Hopping Hoop, *Amer. Math. Monthly* **104** (1997) 152–154.
2. Béla Bollabás, editor, *Littlewood's Miscellany*, Cambridge University Press, London, UK, 1986, p. 37.
3. Raymond Serway, *Principles of Physics*, Harcourt Brace, Orlando, FL, 1994, p. 118ff.

TIMOTHY PRITCHETT graduated with distinction from the Integrated Science Program at Northwestern University. He subsequently attended the Georg-August-Universität Göttingen, where he received a vordiplom in mathematics, and the University of California Berkeley, from which he holds M.A. and Ph.D. degrees in physics. An Associate Professor at West Point, he divides his time between teaching physics to cadets and pursuing his own research interests in nonlinear optics.
Department of Physics, United States Military Academy, West Point, New York 10996
ht1187@exmail.usma.army.edu

Why Dickson Left Quadratic Reciprocity Out of His *History of the Theory of Numbers*

Della Dumbaugh Fenster

In a 1993 letter to the *Notices of the American Mathematical Society*, Irving Kaplansky called attention to an astonishing omission in the history of mathematics [**27**]. "Everybody knows," Kaplansky asserted, "that Dickson's *History of the Theory of Numbers* covers all of number theory up to about 1918. Right?" "Wrong," he answered solidly, "[t]ry looking up quadratic reciprocity." Kaplansky *is* right. Leonard Eugene Dickson's monumental compendium of the history of number theory excludes the history of quadratic reciprocity, the "crown jewel of elementary number theory" [**27**]. Why? Why did Dickson leave this celebrated number theoretic result out of his *History*? Kaplansky offers a brief explanation: "he farmed it out to a student" [**27**]. Again, Kaplansky is right, on some level at least.

In this paper, we offer further insight into this perplexing omission. In the process, we reveal an entirely new perspective on Dickson and unfold yet another example in the history of mathematics where extra-mathematical factors contribute to the development of mathematics. The history of Dickson's *History* actually begins in the last decade of the nineteenth century.

While Dickson pursued a Ph.D. at the young Chicago from 1894 to 1896, the then group-theoretically minded E. H. Moore inspired him to write a thesis on (what we would call) permutation groups [**15**]. Although group theory would remain among Dickson's research interests throughout his career, he would add finite field theory, invariant theory, the theory of algebras, and number theory to his repertoire of research interests. In the spring of 1900, just a few months past his twenty-sixth birthday, the Chicago Mathematics Department invited Dickson to join them as an assistant professor. From this position, Dickson made significant contributions to the consolidation and growth of the algebraic tradition in America [**23**]. Specifically, Dickson spent forty years (all but the first two) of his professional career on the faculty at Chicago where he directed 67 Ph.D. students, wrote 18 books and roughly 300 manuscripts, served as editor of the *American Mathematical Monthly* and the *Transactions of the American Mathematical Society*, and guided the American Mathematical Society as its President from 1916 to 1918 [**3**].

Yet, this mathematical workhorse, who played billiards and bridge by day and did mathematics from 8:30 to 1:30 a.m. every night [**1**, 377], interrupted his thriving pure mathematical career for nearly a decade to write a three-volume, 1500 page historical account of the theory of numbers. As he explained it himself, he undertook this project because "it fitted with my conviction that every person should aim to perform at some time in his life some serious useful work for which it is highly improbable that there will be any reward whatever other than his satisfaction therefrom" [**17**, 2:xxi]. Although he viewed it as "highly improbable," this altruistic mission paid handsome rewards for Dickson as this historical study ultimately led to his celebrated work in the arithmetics of algebras [**23**].

Dickson's description of this historical undertaking as "serious useful work," however, proved more than accurate. This was no hastily written history of number theory. On the contrary, Dickson had planned both the content of his project and

the precise method he would follow to present the details of his study. He revealed the scope of his plans when he explicitly stated his bold intention to "give an adequate account of the entire literature of the theory of numbers" [**17**, 1:iii]. As for his method, the following excerpt from Dickson's *History* reveals both the thoroughness of his study and the historiographic view he maintained throughout this work. On the development of the theory of perfect numbers, he included, for example, that

> Hrotsvitha, a nun in Saxony, in the second half of the tenth century, mentioned the perfect numbers 6, 28, 496, 8128.
>
> Abraham Ibn Ezra (1167), in his commentary to the Pentateuch, Ex. 3, 15, stated that there is only one perfect number between any two successive powers of 10.
>
> Rabbi Josef b. Jehuda Ankin, at the end of the twelfth century, recommended the study of perfect numbers in the program of education laid out in his book "Healing of Souls."
>
> Jordanus Nemorarius (1236) stated (in Book VII, props. 55, 56) that every multiple of a perfect or abundant number is abundant, and every divisor of a perfect number is deficient. He attempted to prove (VII, 57) the erroneous statement that all abundant numbers are even.
>
> Leonardo Pisano, or Fibonacci, cited in his Liber Abbaci of 1202, revised about 1228, the perfect numbers
>
> $$\tfrac{1}{2}2^2(2^2-1)=6,\quad \tfrac{1}{2}2^3(2^3-1)=28,\quad \tfrac{1}{2}2^5(2^5-1)=496$$
>
> excluding the exponent 4 since 2^4-1 is not prime. He stated that by proceeding so, you can find an infinitude of perfect numbers [**17**, 1: 5].

In 1500 pages, Dickson never swerved from this comprehensive, facts-only style of writing. This strict style, in the opinion of the number theorist D. N. Lehmer, made "the book... not so much a history as a list of references from which a history of the theory of numbers might be written" [**28**, 131–132].

To be sure, the reviews of this historical text indicate that Dickson made minor errors in his account. The operative word here is minor—he did not omit major contributions to number theory from his compendium—save the one under discussion. In fact, the reviews of this masterpiece suggest that Dickson accomplished his historical endeavor with the same prowess as his work in pure mathematics. As Robert Carmichael, a number theorist who read the proof sheets for the entire second and third volumes, expressed it in his review for this MONTHLY,

> To give an adequate account of the entire literature of so vast a subject and one of such long history as the theory of numbers is an undertaking of enormous magnitude; and it is carried through in this work with a marvelous success in the presence of which one must pause in admiration. Henceforth this history will be indispensable to all investigators in the theory of numbers... It is a piece of work for which one cannot find a parallel in the whole of scientific history [**5**, 397; 403].

Dickson's *History* remains the classic reference on number theory up to 1918. It provided—and provides?—an "indispensable" source for those lacking adequate library facilities [**5**, 397]. In particular, as Dickson intended, the many "amateurs" interested in mathematics benefited from this (reputably) comprehensive, available

account of number theory [**17**, II: xx] and [**5**, 397].[1] As for the professional mathematician, Lehmer emphasized "the greatest need for just such a piece of work to promote efficiency among the professional workers in this field and to prevent them from wasting their time on problems that have already been adequately treated, and also to suggest other problems which still defy analysis" [**28**, 132]. Lehmer made this point in his review of volume I of Dickson's *History*. The research mathematician would gain much more than "efficiency" by the time all three volumes appeared in print.

Dickson's "systematic" study of Diophantine Analysis for the second volume of his *History*, for example, provided him with a unique, sweeping perspective on this area of mathematics. From this vantage point, Dickson could assert that "[s]ince there already exist too many papers on Diophantine Analysis which give only special solutions, it is hoped that all devotees of this subject will in future refrain from publication until they obtain general theorems on the problem attacked if not a complete solution of it. Only in this way will the subject be able to retain its proper position by the side of other virile branches of mathematics" [**17**, 2: xx]. Dickson, in no uncertain terms, made this assertion with authority. Who better than a prominent research mathematician studying the "disjointed elements" of Diophantine Analysis, could so confidently declare in essence that "[i]deas rather than computations are needed in this field"?[2] Dickson's firm grasp on the past allowed him to see what would lead to a prosperous future for Diophantine Analysis.[3] Interestingly, he himself would devote the final fifteen years of his mathematical career focused on establishing a general result in Diophantine Analysis. But we have gotten ahead of our story. Simply put, Dickson made the history of number theory work in very utilitarian ways—far beyond serving solely as a reference volume—for the research mathematician.[4]

Even still, Dickson's purportedly complete history of the theory of numbers lacks the quintessential topic of elementary number theory, the law of quadratic reciprocity. This law relates the solvability of the congruences $x^2 \equiv p \pmod{q}$ and $x^2 \equiv q \pmod{p}$ for p and q distinct, odd primes. Specifically, if p or q is of the form $4k + 1$ (for $k \in \mathbf{Z}$), the two congruences are both solvable or both not solvable. If p and q are both of the form $4k + 3$ (for $k \in \mathbf{Z}$), one of the congruences is solvable and the other is not. In terms of the Legendre symbol, for

[1] In the letter to President R. S. Woodward of the Carnegie Institution, where he first put forth the idea of a History of Number Theory [**13**], Dickson described his "aim to make a volume indispensable to the specialists, but also a magnet to draw hold the attention of those non-specialists who desire to secure a connected scientific account of the subject." This theme of appealing to professional and amateur alike appears throughout Dickson's correspondence with the Carnegie Institution regarding his *History*.

[2] Both quotations are from [**6**, 72–73]. This emphasis on general results could easily be viewed as a "Dicksonian trademark." In his development of the definition of an algebra, Dickson sought the definition that yielded the theory with the widest applications. Similarly, in his work on the arithmetic of algebras, Dickson built his definition of an integral element, the crucial concept in the theory, using a "strategy of enlargement" as employed by Kummer and Gauss rather than defining an integral element on a case-by-case basis. See [**22**] and [**23**, 139–143; 152].

[3] In [**7**, 262], Carmichael emphasized the value of such forecasting when he wrote that "[w]hen a master, with the work of the past well in mind, tries to see the trend of the future, his judgement will be a matter of interest whether or not the direction of progress turns out to be such as he anticipates. It may even throw some light on the difficult question as to the way in which new discoveries arise."

[4] Dickson's *History* inspired—in the most general sense of the word—one prominent twentieth-century mathematician. Richard Guy purchased Dickson's *History* when he was about seventeen and found it "better than getting the whole works of Shakespeare and heaven knows what else" [**2**, 136].

p and q distinct, odd primes,

$$\left(\frac{p}{q}\right)\left(\frac{q}{p}\right) = (-1)^{\left(\frac{p-1}{2}\right)\left(\frac{q-1}{2}\right)}.$$

This law, as Dickson described it himself, "is doubtless the most important tool in the theory of numbers and occupies the central position in its history. Its generalizations form a leading topic, past and present, in the theory of algebraic numbers" [**18**, 30]. Since the development of algebraic number theory grew, in large part, out of efforts to generalize quadratic reciprocity, it seems all the more unusual that a supposedly comprehensive *History of the Theory of Numbers* included no discussion of this area.[5]

Why, then, did Dickson exclude an account of "this most important" tool from his *History*? The historical record suggests that Dickson did not intend for this omission to occur. In his closing remarks in the preface to volume II, Dickson refers to a Volume III as the "concluding" volume in the series [**17**, 2: xii]. Volume III appeared in 1923, "promptly" prepared, as Dickson described it in the preface, "owing to the favorable reception accorded to the first two volumes of this history" [**17**, 3: iii]. Early in the text of this third volume, nestled in his history of binary quadratic forms, Dickson points us forward to a *fourth* volume [**17**, 3: 3]. In this parenthetical remark, Dickson indicated his plan to include the quadratic reciprocity law in the fourth volume. But, of course, as we know now, the fourth volume never appeared. What happened to the fourth volume?

The fourth volume involves Albert Everett Cooper, a University of Chicago graduate student from 1924–1926. Cooper had come to Chicago from the University of Texas where he had earned three degrees and taught as an instructor in the mathematics department for 5 years [**9**, vita following dissertation text]. Arriving as he did in the fall of 1924, Cooper met a Dickson who had just spent over a decade collecting references on number theory, "digesting" them, and writing them up in a suitable historical account of the subject, which ultimately grew from one to two to three volumes and spanned some 1500 pages [**14**]. This same Dickson had recently managed to secure publication for volume III of his *History* despite a change in presidency at the Carnegie Institution. He had just published his work in the arithmetics of algebra. He had now begun to focus his research interests on number theoretic topics, increasingly inclined toward problems related to extensions of Waring's Theorem. Perhaps more to the point, Cooper met a Dickson who had promised the Carnegie Institution, the mathematical world, and himself a comprehensive account of number theory, and, thus far, had failed to include information on the history of quadratic residues or reciprocity laws.

Cooper earned his Ph.D. in mathematics in the spring of 1926 under Dickson's guidance, with his historical dissertation "A Topical History of the Theory of Quadratic Residues" [**9**]. The title of Cooper's dissertation alone suggests a connection with Dickson's larger historical undertaking. Indeed, Cooper wrote this

[5] In [**16**], Dickson claimed that "the study of this challenge problem [Fermat's Last Theorem] and the general law of reciprocity of higher residues led [Ernst Eduard] Kummer to invent his ideal numbers, out of which grew the general theory of algebraic numbers, one of the most important branches of modern mathematics [p. 161]." This article seems to represent an expanded version of a similar discussion in his *History*, 2, pp. xviii–xix, 739–740. In [**21**, 324], Harold M. Edwards advances the view that the study of higher reciprocity laws and not Fermat's Last Theorem led Kummer to his study of what we know as cyclotomic integers and, ultimately, his ideal numbers. Edwards developed this idea more thoroughly in [**20**, 79–81].

dissertation with the intention that it appear as a chapter in the fourth volume of Dickson's history. Cooper cited earlier volumes of "this History," as he referred to it, more than twenty times in his 98-page thesis. In Cooper's section on "Number and Distribution of Residues," for example, Cooper asserted that "Dirichlet's fundamental formula for the distribution, in half, quarter, and eighth intervals, of the quadratic residues of a positive odd integer P stated in terms of the class number of a binary quadratic form of negative determinant are *quoted in vol. III, page 101 of this History*" [**9**, 53–53A, our emphasis].

More importantly, at least relative to the study at hand, Cooper's dissertation did not include a history of the law of quadratic reciprocity. Cooper did not overlook this "principal" theorem, however. Cooper titled his dissertation appropriately; he wrote a historical account of the theory of quadratic residues. On at least sixteen occasions in his dissertation, in fact, Cooper referred to a specific chapter on quadratic reciprocity in the purported fourth volume. "For proof [of G. Zolotareff's "unique method" of evaluation of $\left(\frac{k}{p}\right)$ where $k, p \in \mathbf{Z}$, p a prime, and p does not divide k]," Cooper wrote in his dissertation for example, "see chapter on quadratic reciprocity, this History, vol. IV" [**9**, 35]. He ended the text of his dissertation pointing to the "chapter on the law of quadratic reciprocity" and leaving spaces to assign both this chapter a number and the results of eight mathematicians page numbers within this chapter. Thus, Cooper wrote his dissertation to form a chapter on quadratic residues in Dickson's fourth volume, perhaps in the same spirit as G. H. Cresse's chapter on the class number in volume III.[6] Dickson's fourth volume, as Cooper understood it, would contain a separate chapter on the history of quadratic reciprocity. Dickson must have had this same understanding; he, after all, approved Cooper's thesis.

Dickson did more than simply approve Cooper's thesis at the end; he also gave him a boost at the beginning. Cooper described Dickson as an advisor who "not only furnished the entire body of original references from which the topical history was written, but also took a great deal of personal interest in the preparation of the material. My appreciation is particularly due Professor Dickson" [**9**, vita, immediately following dissertation text]. Although Cooper probably intended that this comment serve as an acknowledgement to his thesis advisor, this remark indicates that Dickson had amassed a collection of references on the history of quadratic residues by the time Cooper arrived at Chicago in 1924. H. S. Vandiver further substantiates this claim in a letter he wrote to Cooper on 5 November, 1925. "I'm sending you under separate cover," Vandiver wrote in response to Cooper's request for copies of his articles relating to quadratic residues, "copies of my articles which touch on quad-residues . . . . Perhaps you know that Dickson collected references of quad. residues while writing his *History*" [**32**]. Since Dickson collected quadratic residue references, surely he must have intended to include them, in some way at least, in his *History*. Although it may not have been his initial plan when he conceived of his *History* in 1911, Dickson ultimately entrusted the history of quadratic residues to a graduate student, namely, Albert Cooper.

[6]George Hoffman Cresse earned a Chicago Ph.D. in 1918 under Dickson's guidance with the historical dissertation "On the Class Numbers of Binary Quadratic Forms" [**30**]. A revised form of this dissertation appeared as chapter VI in the third volume of Dickson's *History*. The official title page of the third volume of Dickson's *History* reads, "History of the Theory of Numbers, Volume III, Quadratic and Higher Forms, By Leonard Eugene Dickson, Professor of Mathematics at the University of Chicago, With a Chapter on the Class Number, by G. H. Cresse."

Perhaps the more compelling question raised by this evidence surrounds the publication of Cooper's work. What happened to Cooper's thesis work on quadratic residues? And where is the chapter on quadratic reciprocity?

After Cooper completed his Ph.D. at Chicago, he rejoined the University of Texas mathematics faculty. A year later, in 1927, Dr. Harry Y. Benedict, a former classmate from Dickson's undergraduate days at Texas, assumed the presidency of that institution. Dickson gave his view on the election of Benedict to the presidency of the University of Texas in the alumni publication, *The Alcalde*. As Dickson saw it, "[t]he election of Dr. Benedict as President of the University of Texas is particularly fortunate. All are familiar with his success as dean, due to his unerring judgement, rare talents as an executive, and deep affection for the University. But I wish to emphasize the fact that the man having all these essential qualities is also a scientist [astronomer]. This is the age of science..." [**4**] (and partly quoted in [**29**, 233]). Within months of penning this favorable response to Benedict's election, Dickson personally appealed to Benedict to finance a mathematical publication. That publication? The fourth and final volume of the *History of the Theory of Numbers*.

The letter to Benedict came on the heels of a two-month-long exchange of correspondence between Dickson and the Carnegie Institution regarding the publication of the fourth volume of Dickson's *History*. On 1 December, 1927, Dickson wrote the Carnegie Institution regarding the recently expressed interest of G. E. Stechert & Co. to reprint volumes I and II of Dickson's *History*. Dickson advised "allowing Stechert to reprint" and explained that "[w]e think that even if Carnegie Inst[itutio]n could afford to reprint, it would be wiser for it to spend same sum to print (short) vol IV of the History and so complete the series and to spend the balance on a History of the Solution of *Equations*" [**11**]. The administrative secretary of the Carnegie Institution, W. M. Gilbert, replied that the Institution would both prefer Stechert to reprint the two volumes and "be glad to have an opportunity to issue your fourth and final volume..." [**25**].

Dickson, however, had more on his mind than reprinting his *History* and publishing his fourth volume. Apparently, since the first of December he had outlined a new plan in a series of letters to Gilbert and John C. Merriam, the president of the Carnegie Institution. "My suggestion is as follows," Dickson wrote to Gilbert and Merriam at the end of that December,

> [l]et the Institution abandon not only the reprinting of Vols I & II, but also the printing of Vol. IV. Instead, let the Institution take on half the burden of publishing an entirely new work on the Theory of Numbers (which will meet all legitimate needs of a History, but also attend to the more important needs of presenting the whole theory of numbers as a science, with emphasis on methods). As I wrote before, the University of Chicago Press is committed to publishing *one* of the two necessary volumes....
>
> The new work will be incomparably superior to the old History; will be what is needed permanently in this field; and will be a fitting sequel to my 35 years research in this field [**12**].

Thus Dickson himself proposed that the Carnegie Institution no longer plan to publish the final, fourth volume of his *History*, but rather, agree to publish one of two new forthcoming volumes on Number Theory. Dickson felt that "an abbreviated Vol. IV might be published by some agency other than the Institution (& I would so undertake)" [**12**]. Apparently, over the course of the next month,

Dickson determined that his old classmate Benedict and the University of Texas might just prove the best possible path of publication for volume IV.

"My dear Benedict," Dickson began in his letter of 27 January, 1928, "Dr. Cooper has now practically complete his historical ms on 'Quadratic Residues and Reciprocity Law.' He has done a fine job" [**10**]. This opening sentence indicates that between the spring of 1926, when Cooper earned his Ph.D. from Chicago, and January of 1928, Cooper had written "the chapter" on the history of quadratic reciprocity. Dickson explained to Benedict that he had originally planned to include Cooper's historical manuscript on quadratic residues and reciprocity law in the final, fourth volume of his *History*. If he carried out this plan, however, he would have to include a supplement containing corrections and additions to the first three volumes. (In [**17**, 3: iv], Dickson asked readers to send "errata or omissions, which will be published later as a supplement.") "There are," as Dickson tersely explained it, "reasons against my undertaking [a supplement]." "Also," Dickson added, now shifting the focus from himself to Cooper, "the objection that such a vol. IV would not be wholly the work of Cooper" [**10**]. Although Dickson did not cite the source of this "objection," he seemed to imply that Cooper (and, presumably, the University of Texas by association?) would not receive the credit he deserved if his work appeared as chapters in Dickson's fourth volume. It would be better for Cooper, Dickson emphasized, if he published his manuscript separately, perhaps, as Dickson suggested, with a subtitle to indicate that it formed the fourth volume of his *History*.

Dickson told Benedict that the "problem," as he referred to it, had further complications. Specifically, by this time, the first two volumes were out of print. Dickson had planned to "replenish" the information in volumes I & II by writing a 2-volume historical and expository account of the Theory of Numbers. As Dickson presented it to Benedict, the University of Chicago agreed to print one of these volumes and the Carnegie Institution agreed to print the other. There was a condition, however. The Carnegie Institution would print one of these two new historical and expository accounts of number theory only if relieved of any responsibility for printing a fourth volume in the original historical series. It seems Dickson gently twisted the details to attempt to secure publication for both his two new works on Number Theory and "his" volume IV.

Having spelled out all of these details, Dickson proposed "the following best plan to serve the interests of Mr. Cooper and mathematics [and himself?]: Let the University of Texas publish a book by Dr. Cooper on the History of Quadr[atic] Residues and Reciprocity Law. This would close up the gap now existing from lack of vol. IV of my History and would take the place of the latter" [**10**]. Although Dickson promised sales for a book by Cooper and urged Benedict to follow the lead of other large universities who aided in the publication of serious work done by their faculties, we know Benedict did not agree to publish this work of Cooper's as the subtitled fourth volume in the series. We know Benedict did not agree to publish the fourth volume, but we do not know why. His reply to Dickson remains lost.[7] Benedict replied to hundreds of requests, including those of garden clubs

[7]Here, by lost we mean that the copy of the letter Benedict sent Dickson is not in any of the "natural" or even some of the "unnatural" collections in the Texas Archives. Of course, Benedict could have responded to Dickson by phone. As Albert Lewis points out in [**29**, 207], in the 1920's university presidents began responding to more sensitive and controversial issues by telephone. In these cases, the president usually noted "Answered by telephone" at the bottom of the original letter. Dickson's letter bears no such annotation.

and small girls schools in south Texas; he must have replied to this letter. Since Dickson burned his papers upon his retirement, he probably destroyed the original copy of Benedict's response.

After January of 1928, Dickson apparently never mentioned volume IV in his correspondence with the Carnegie Institution again. Cooper, however, sent a telegram to the Institution in June of 1929 announcing that "I have ready the manuscript for the fourth volume." He queried further, "How many copies do you think should be printed?" [**8**] The secretary of the Division of Publications of the Carnegie Institution sent Cooper his advice but added that "[i]t seems to me that, with the information given here, Dr. Dickson would be the best one to decide this matter" [**26**]. Although Cooper and Dickson exchanged scores of papers, notes, and communiques of various forms on the history of quadratic reciprocity, they did not seem to leave any traces of a discussion of the publication of this material [**31**].

The organized and polished pieces of this collection, however, appear to represent the page proofs of a book written in the same spirit and style of the first three volumes of Dickson's *History*. (Perhaps these form the fourth volume Cooper referred to in his telegram?) The more loose and ordinary fragments (written on "scrap" paper, scrawled on the back of notices from the registrar, etc.) seem to provide brief summaries of and references to various articles on quadratic residues and reciprocity. The collection contains notes and papers written by both Cooper and Dickson. Cooper, however, maintained the collection. It seems unlikely that Dickson would have continued to supply Cooper with the information if he intended to write this portion of the history himself. Moreover, as Dickson indicated in his letter to President Benedict, Dickson could not have written the main section on the history of quadratic residues because Cooper had already done it. So either Dickson intended to publish the history as a collaborative effort with Cooper or, more likely, he saw Cooper as publishing it himself.

The question now becomes "why did Dickson not forge ahead with the fourth volume in some form?" It was totally uncharacteristic of Dickson to leave his work incomplete, as it were. He had a plan, at least in late 1927 and early 1928, to see this material to press. Cooper still persisted with ideas of publication as late as June of 1929.[8] But the work never appeared. Perhaps, by the late 1920's, with his historical project more than fifteen years old and his research program devoted almost exclusively to Waring's Problem, Dickson found himself completely occupied with other mathematical endeavors. Maybe his interest waned in the historical text, maybe he mentally turned over the fourth volume and its publication to Cooper, or, maybe, his "astonishing supply of energy" finally evaporated.[9]

Whatever the case, our historical study leads to an intriguing observation. Leonard Dickson sits squarely in the center of this episode in the history of (American) mathematics. Yes, Leonard Dickson, the prolific mathematician who had a reputation for completeness, for high standards, for excellence—even to the point of being impolite when insisting upon these standards [**24**, 13–14]. And, yet, this same Dickson had to come to terms with the non-appearance of this signifi-

[8]Aside from the aforementioned telegram from Cooper to the Carnegie Institution in June of 1929, we have no other record of Cooper's attempt(s?) to publish this material.

[9]Interestingly, in [7, 259], Carmichael subtly hinted at the possibility of Dickson running out of steam when he wrote that "[t]he reviewer ventures to predict that the favorable reception of the third volume will give the author still more reason for proceeding promptly with the fourth if his astonishing supply of energy is holding out well enough to leave him still susceptible to such influence."

cant result in number theory. This set of events in the history of mathematics certainly sheds new light on Dickson the mathematician and the man.

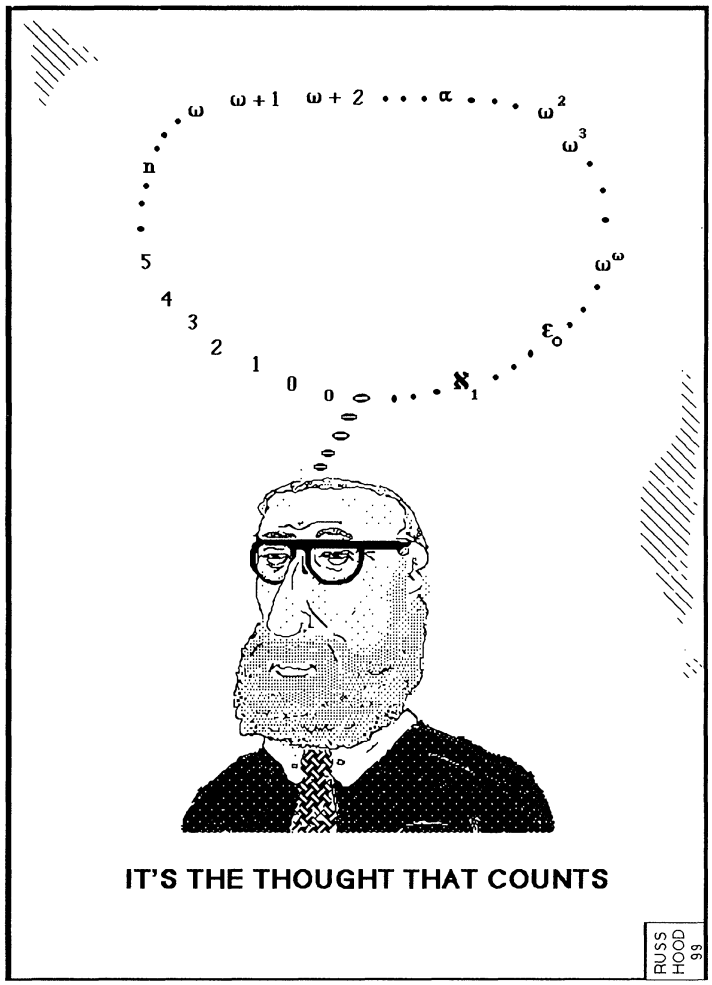
In the end, then, once again, the history of mathematics teaches us that mathematics—and mathematicians, for that matter—are more than they appear. In particular, mathematical and extra-mathematical factors impinge upon the development and publication of mathematics, and the history thereof. In this case, clearly, the extra-mathematical factors, in the form of authorship priority, publishing contracts, and finances, outdistanced the mathematical factors and resulted in a highly unusual omission.

REFERENCES

1. Donald J. Albers and G. L. Alexanderson, A Conversation with Ivan Niven, *College Math. J.* **22** (1991) 307–402.
2. Donald J. Albers and G. L. Alexanderson, A Conversation with Richard K. Guy, *College Math. J.* **24** (1993) 123–148.
3. A. Adrian Albert, Leonard Eugene Dickson 1874–1954, *Bull. Amer. Math. Soc.* **61** (1955) 331–345.
4. *The Alcalde*, 27 November, 1927, The Center for American History, The University of Texas at Austin.
5. Robert D. Carmichael, Review of *History of the Theory of Numbers*, Vol. I: Divisibility and Primality, *Amer. Math Monthly* **26** (1919) 396–403.
6. Robert D. Carmichael, Review of *History of the Theory of Numbers*, Vol. II: Diophantine Analysis, *Amer. Math. Monthly* **28** (1921) 72–78.
7. Robert D. Carmichael, Review of *History of the Theory of Numbers*, Vol. III: Quadratic and Higher Forms, *Amer. Math. Monthly* **30** (1923) 259–262.
8. A. E. Cooper to Carnegie Institution, 25 June, 1929, Carnegie Institution Archives, Washington D. C., Dickson Papers.
9. Albert Evertt Cooper, "A Topical History of the Theory of Quadratic Residues," University of Chicago Archives.
10. Leonard Dickson to Harry Y. Benedict, 27 January, 1928, H. S. Vandiver Papers, Box 12, Folder 1, The Center for American History, The University of Texas at Austin.
11. Leonard Dickson to Carnegie Institution, 1 December, 1927, Carnegie Institution Archives, Washington D. C., Dickson Papers.
12. Leonard Dickson to W. M. Gilbert and John C. Merriam, 29 December, 1927, Carnegie Institution Archives, Washington D. C., Dickson Papers.
13. Leonard Dickson to R. S. Woodward, 11 February, 1911, Carnegie Institution Archives, Washington D. C., Dickson Papers.
14. Leonard Dickson to R. S. Woodward, 23 October, 1911, Carnegie Institution Archives, Washington D. C., Dickson Papers.
15. Leonard Dickson, The Analytic Representation of Substitutions on a Power of a Prime Number of Letters with a Discussion of the Linear Group, *Ann. of Math.* (2) **11** (1897) 65–143.
16. _____, Fermat's Last Theorem and the Origin and Nature of the Theory of Algebraic Numbers, *Ann. of Math.* (2) **18** (1917) 161–187.
17. _____, *History of the Theory of Numbers*, 3 vols. New York: Chelsea Publishing Company, 1919, 1920, 1923.
18. _____, *Introduction to the Theory of Numbers*, Chicago: The University of Chicago Press, 1929.
19. _____, *Studies in the Theory of Numbers*, Chicago: The University of Chicago Press (The University of Chicago Science Series), 1930.
20. Harold M. Edwards, *Fermat's Last Theorem*, Berlin: Springer, 1977.
21. _____, The Genesis of Ideal Theory, *Arch. Hist. of Exact Sci.* **23** (1980) 321–378.
22. Della D. Fenster, The Development of the Concept of an Algebra: Dickson's Role, forthcoming.
23. _____, Leonard Eugene Dickson and his work in the Arithmetics of Algebras, *Arch. Hist. Exact Sci.* **52** (1998) 119–159.
24. _____, Role Modeling in Mathematics: The Case of Leonard Eugene Dickson (1874–1954), *Historia Math.* **24** (1997) 7–24.
25. W. M. Gilbert to Leonard E. Dickson, 20 December, 1927, Carnegie Institution Archives, Washington D. C., Dickson Papers.

26. Irving M. Grey to A. E. Cooper, 26 June, 1929, Carnegie Institution Archives, Washington D. C., Dickson Papers.
27. Irving Kaplansky, Quadratic Reciprocity in Dickson's *History*, *Notices Amer. Math. Soc.* **40** (1993) 1155.
28. D. N. Lehmer, Dickson's History of the Theory of Numbers, *Bull. Amer. Math. Soc.* **26** (1919) 125–132.
29. Albert C. Lewis, "The Building of the University of Texas Mathematics Faculty," in *A Century of Mathematics in America—Part III*, ed. Peter Duren, Providence: American Mathematical Society, 1989, pp. 205–239.
30. University of Chicago, Department of Special Collections, "Doctors of Philosophy, June, 1893–April 1931," in *Announcements: The University of Chicago*, vol. 31, number 19, May 15, 1931, Chicago: Chicago Press, 1931.
31. A. E. Cooper Papers, The Center for American History, The University of Texas at Austin.
32. H. S. Vandiver to A. E. Cooper, 5 November, 1925, A. E. Cooper Papers, Box 1, Folder 2, The Center for American History, The University of Texas at Austin.

DELLA FENSTER is an assistant professor of mathematics at the University of Richmond. Her research focuses primarily on the history of American Mathematics (especially algebra and number theory) in the late nineteenth and early twentieth centuries.
Department of Mathematics and Computer Science, University of Richmond, Richmond, Virginia, 23173
dfenster@richmond.edu



Contributed by Russ Hood, Rio Linda, CA

Uniform Calculus and the Law of Bounded Change

Mark Bridger and Gabriel Stolzenberg

1. INTRODUCTION. In a recent exchange about the role of the mean value theorem in the theory of the calculus, T. Tucker notes that "the origin of the Mean Value Theorem in the structure of the real numbers" is much too difficult for a standard course [6]. He shows how the increasing function theorem (a function with positive derivative is increasing) serves very nicely in place of the mean value theorem, and sketches a proof of it from the nested interval property of the real number system.

In support of the mean value theorem, H. Swann recalls its derivation from the extreme value theorem (a continuous function on a closed interval has a maximum value) via Rolle's theorem and remarks that "such a sequence of arguments reveals the charm and power of mathematics, for we prove that a questionable complicated result *must* be true if we assume other simpler results that are less questionable" [5].

We agree with Swann about the charm and power of mathematics and with Tucker about the ability of the increasing function theorem to play a role traditionally accorded the mean value theorem. In fact, we give several examples that support Tucker's claim. But Tucker and Swann work with *pointwise* continuity and differentiability, weak notions that make proving statements like the increasing function theorem more difficult. On closed finite intervals, uniform continuity and differentiability are as easy to verify, and using them as starting points permits a natural development of the calculus in which such difficulties do not arise.

Our treatment of continuity and differentiation is from our forthcoming book, *A New Course of Analysis*, where it is expressed in terms of a theory of real numbers based on interval order and arithmetic. We offer no theory of real numbers in this article but we use repeatedly the fact that each real number can be approximated by rationals to arbitrary accuracy.

2. UNIFORM CALCULUS. Continuity. Uniform continuity of a one variable function f is a condition on its variation, $f(y) - f(x)$. The condition, written $\phi(x, y) \to 0$ as $y - x \to 0$ for any two-variable function ϕ, is that for each $\epsilon > 0$, there is a $\delta > 0$ such that $|\phi(x, y)| \leq \epsilon$ if $|y - x| \leq \delta$. When $\phi(x, y) = U(x, y) - u(x)$, we also write: $U(x, y) \to u(x)$ as $y \to x$.

Example. The relationship

$$y^n - x^n = \left(\sum_{i=1}^{n} y^{n-i} x^{i-1} \right)(y - x)$$

shows that $|y^n - x^n| \leq nC^{n-1}|y - x|$ on any interval of the form $[-C, C]$ and, hence, that x^n is uniformly continuous on each finite interval. A proof of pointwise continuity could hardly be simpler.

Example. Using $p^n + q^n \le (p+q)^n$ with $p = x^{1/n}$ and $q = (y-x)^{1/n}$, we have

$$0 \le y^{1/n} - x^{1/n} \le (y-x)^{1/n} \le \epsilon \quad \text{if } 0 \le x \le y \quad \text{and } y - x \le \epsilon^n.$$

It follows that for each positive integer n, $x^{1/n}$ is uniformly continuous on $[0, \infty)$.

Proposition 2.1. *A composition of uniformly continuous functions is uniformly continuous.*

Proposition 2.2. *A uniformly continuous function f on a finite interval I is bounded.*

Proof: For $\epsilon > 0$, let $\delta > 0$ be given by uniform continuity. Because I is finite, we can find finitely many points such that every $p \in I$ is within δ of at least one of them. Hence, f is bounded by ϵ plus the maximum of its values at these finitely many points. ∎

Differentiability. Uniform differentiability of a function f also is a condition on its variation: it factors as $f(y) - f(x) = F(x, y)(y - x)$, where $F(x, y) \to F(x, x)$ as $y \to x$. If f is uniformly differentiable, its derivative is the function $f'(x) = F(x, x)$. Thus, $F(x, y) = (f(y) - f(x))/(y - x)$, for y different from x, and $F(x, x) = f'(x)$.

Because the difference quotient converges to the derivative as $y \to x$, the derivative is unique on any domain S for which each x in S is approximable to arbitrary accuracy by points y in S different from x.

Example. For all positive integers n, using the factorization of $y^n - x^n$ and the arithmetic of convergence (see Lemma 4.1), it follows that on each finite interval, x^n is uniformly differentiable with derivative nx^{n-1}.

Example. Because $y^2 - x^2 = (y + x)(y - x)$ and $y + x \to 2x$ on $\mathbb{R}$ as $y \to x$, x^2 is uniformly differentiable on $\mathbb{R}$ with derivative $2x$.

Proposition 2.3. *If f is uniformly differentiable, then f' is uniformly continuous.*

Proof: Because F is symmetric, if x and y are close enough, both $f'(x)$ and $f'(y)$ are within ϵ of $F(x, y) = F(y, x)$ and hence within 2ϵ of each other. ∎

Corollary 2.4. *On finite intervals, f' is bounded.*

Proof: Propositions 2.2 and 2.3. ∎

Proposition 2.5. *If f' is bounded, then f is uniformly continuous.*

Proof: When f' is bounded, so is $F(x, y)$, say by C, for $|y - x|$ sufficiently small. Hence, $|f(y) - f(x)| \le C|y - x| \to 0$ as $y - x \to 0$. ∎

Theorem 2.6. (*Fundamental Theorem of the Calculus*) *If g is uniformly continuous on $[a, b]$, then $G(x) \equiv \int_a^x g(t)\,dt$ is uniformly differentiable on $[a, b]$ with $G' = g$.*

Proof: $G(y) - G(x)$ equals the integral of g from x to y, which equals $y - x$ times a limit of averages of values of g at points in $[x, y]$. (To see this, approximate the integral by Riemann sums with equal spacing.) Also, for each $\epsilon > 0$, if $|y - x|$

is small enough, every value of g at a point in $[x, y]$ is within ϵ of $g(x)$. But then, also, any limit of averages of values of g at points in $[x, y]$ is within ϵ of $g(x)$, so we are done.

3. THE ARITHMETIC OF UNIFORM CONTINUITY. The arithmetic of uniform continuity is very simple. If both f and g are uniformly continuous, so is $f + g$. If also f and g are bounded, then fg is uniformly continuous. Finally, if $1/f$ is defined and bounded, it too is uniformly continuous.

These statements can be verified by first relating the variations of the sum, product and reciprocal to those f and g. Simple algebra shows that $\mathrm{var}(f + g) = \mathrm{var}(f) + \mathrm{var}(g)$, $\mathrm{var}(fg) = g(y)\mathrm{var}(f) + f(x)\mathrm{var}(g)$ and $\mathrm{var}(1/f) = -\mathrm{var}(f)/(f(x)f(y))$.

Because each expression is a sum of expressions of the form $B(x, y)\phi(x, y)$, where $B(x, y)$ is bounded and $\phi(x, y) \to 0$ as $y - x \to 0$, it suffices to verify that each such sum again converges to 0 as $y - x \to 0$. We omit the simple proof of this.

4. THE ARITHMETIC OF UNIFORM DIFFERENTIABILITY. If f and g are uniformly differentiable, derivatives for their arithmetic combinations are given by the following rules.

Sums. $f + g$ is uniformly differentiable with $(f + g)' = f' + g'$.

Products. If f, g, and their derivatives are bounded, e.g., if their domain is a finite interval, then fg is uniformly differentiable with $(fg)' = f'g + fg'$.

Reciprocals. If $1/f$ is defined and bounded, and f' also is bounded, then $1/f$ is uniformly differentiable with $(1/f)' = -f'/f^2$.

To prove these assertions, we begin by substituting $F(x, y)(y - x)$ and $G(x, y)(y - x)$ for $\mathrm{var}(f)$ and $\mathrm{var}(g)$ in our expressions for $\mathrm{var}(f + g)$, $\mathrm{var}(fg)$, and $\mathrm{var}(1/f)$. For the sum, we get $F(x, y) + G(x, y)$, for the product, $g(y)F(x, y) + f(x)G(x, y)$, and for the reciprocal, $-F(x, y)/(f(x)f(y))$, each multiplied by $y - x$.

For $y = x$, these expressions become $f'(x) + g'(x)$, $g(x)f'(x) + f(x)g'(x)$, and $-f'(x)/f^2(x)$.

The case of the sum is clear. Both $F(x, y) - f'(x)$ and $G(x, y) - g'(x)$ converge to 0 as $y - x \to 0$, hence so does the sum. For the product and reciprocal, multiplications are involved. The following lemma gives us what we need to deal with them.

Lemma 4.1. *Suppose that u and v are bounded. If $U(x, y) \to u(x)$ and $V(x, y) \to v(x)$ as $y \to x$, then for $y - x$ sufficiently small, U and V are bounded and $U(x, y)V(x, y) \to u(x)v(x)$ as $y \to x$.*

Proof: Write $UV - uv = u(V - v) + (U - u)V$ and note that, because u and V are bounded for $y - x$ small enough, each summand converges to 0 as $y - x \to 0$. ■

The next lemma is used to prove Proposition 6.1 about the differentiability of an inverse function.

Lemma 4.2. *Suppose that* $1/u$ *is defined and bounded. If* $U(x, y) \to u(x)$, *then for* $y - x$ *sufficiently small,* $1/U$ *is defined and bounded, and* $1/U(x, y) \to 1/u(x)$ *as* $y \to x$.

Proof: We prove only the second part. Write $1/u - 1/U = (U - u)/uU$ and note that $1/uU$ is bounded for $y - x$ sufficiently small. ■

For the product rule, we reason as follows. By assumption, f is bounded and $G(x, y) \to g'(x)$ as $y \to x$. Thus, $f(x)G(x, y) \to f(x)g'(x)$ as $y \to x$. Because limits add, it suffices to prove that $g(y)F(x, y) \to g(x)f'(x)$ as $y \to x$. But, also by assumption, $F(x, y) \to f'(x)$ as $y \to x$, and f' and g are bounded. Hence, if $g(y) \to g(x)$ as $y \to x$, we can apply Lemma 4.1. It therefore suffices to note that g is uniformly continuous because g' is bounded.

Similarly, for the reciprocal rule, f is uniformly continuous because f' is bounded, and because $1/f$ is bounded, it too is uniformly continuous. Hence, $1/f(y) \to 1/f(x)$ as $y \to x$. Multiplying by $1/f(x)$, we see that $1/f(x)f(y) \to 1/f^2(x)$ as $y \to x$. Because $-F(x, y) \to -f'(x)$ as $y \to x$, and the limit functions $1/f^2$ and $-f'$ are bounded, the product converges to the product by Lemma 4.1.

5. THE CHAIN RULE

Proposition 5.1. *If* f *and* g *are uniformly differentiable, and if* f' *and* g' *are bounded, then* $f(g)$ *is uniformly differentiable with derivative* $f'(g)g'$.

Proof: Because $f(g(y)) - f(g(x)) = F(g(x),g(y))(g(y) - g(x))$, which in turn equals $F(g(x), g(y))G(x, y)(y - x)$, the candidate for the derivative of $f(g)$ is indeed $f'(g)g'$. Because g' is bounded, g is uniformly continuous. Hence, $F(g(x),g(y)) \to f'(g(x))$ as $y \to x$. Because $G(x, y) \to g'(x)$ as $y \to x$ and $f'(g)$ is bounded, an application of Lemma 4.1 gives us the desired result. ■

6. DIFFERENTIABILITY OF THE INVERSE

Proposition 6.1. *If* h *is a uniformly continuous inverse for* f, *and if* $1/f'$ *is defined and bounded, then* h *is uniformly differentiable with* $h' = 1/f'(h)$.

Proof: Because h is an inverse for f, we can factor the variation of the identity function as

$$y - x = f(h(y)) - f(h(x)) = F(h(x), h(y))(h(y) - h(x)).$$

This shows that $1/F(h(x), h(y))$ is equal to the difference quotient for h when $|y - x| > 0$. Because h is uniformly continuous, $F(h(x), h(y)) \to f'(h(x))$ as $y - x \to 0$. Therefore, because $1/f'(h(x))$ is defined and bounded, we can apply Lemma 4.2 to conclude that $1/F(h(x), h(y)) \to 1/f'(h(x))$ as $y \to x$. ■

7. THE LAW OF BOUNDED CHANGE

Theorem 7.1. *If* f *is uniformly differentiable and* $A \le f' \le B$ *on* $[a, b]$, *then* $A(b - a) \le f(b) - f(a) \le B(b - a)$.

This is the law of bounded change. It says that bounds for the derivative are bounds for the difference quotient. Notice that the increasing function theorem is

just the law of bounded change for $A = 0$ (and we don't care about B) and the law of bounded change is the increasing function theorem applied to the functions $Bx - f(x)$ and $f(x) - Ax$.

Proof: It suffices to prove that for all $\epsilon > 0$, the conclusion holds with A and B replaced by $A - \epsilon$ and $B + \epsilon$. The justification for this is the general truth that if $p < q + \epsilon$ for all $\epsilon > 0$, then $p \le q$. That this holds for reals follows by rational approximation from the fact that it holds for rationals.

Since $F(u, v) \to f'(u)$ as $v \to u$, for each $\epsilon > 0$ there is a $\delta > 0$ such that $f'(u) - \epsilon < F(u, v) < f'(u) + \epsilon$ for $0 \le v - u < \delta$. But $A \le f'(u) \le B$, so $f(v) - f(u) = F(u, v)(v - u)$ lies between $(A - \epsilon)(v - u)$ and $(B + \epsilon)(v - u)$.

Hence, if we express $f(b) - f(a)$ as a telescoping sum of n differences $f(u_i) - f(u_{i-1})$, where $u_0 = a$ and each $u_i - u_{i-1} = (b - a)/n < \delta$, we have that $(A - \epsilon)(b - a) \le f(b) - f(a) \le (B + \epsilon)(b - a)$. ∎

We now draw several useful and easy consequences of the law of bounded change.

Corollary 7.2. *f is constant on any interval on which $f' = 0$.*

Proof: This is just the law of bounded change with A and B equal to 0. ∎

Is there any simpler or essentially different way to prove this deceptively obvious-looking fact?

Corollary 7.3. $f(x) - f(a) = \int_a^x f'(t)\,dt$.

Proof: By the fundamental theorem of the calculus, the two sides of the equation have the same derivative. Hence, by Corollary 7.2, they differ by a constant. But they agree at $x = a$, so they agree everywhere. ∎

Alternatively, we can observe that in the proof of the law of bounded change, we in effect approximate $f(x) - f(a)$ to arbitrary accuracy by Riemann sums for the integral of f' from a to x. Because these sums also approximate the integral, the two must be equal.

Corollary 7.4. *If $f' \ge A > 0$ on $[a, b]$ and $f(h(u)) = u$ for all u in $[f(a), f(b)]$, then h is uniformly continuous.*

Proof: By the law of bounded change, if $h(u) < h(v)$, then $A(h(v) - h(u)) \le f(h(v)) - f(h(u)) = v - u$. So $0 < h(v) - h(u) \le (v - u)/A \to 0$ as $v - u \to 0$. ∎

By the inverse function theorem, whenever $f' \ge A > 0$ on $[a, b]$, there is a function h as in the statement of Corollary 7.4.

Corollary 7.5. *If $A \le f' \le B$ on $[a, b]$, then*

$$\left| \frac{f(y) - f(x)}{y - x} - f'(u) \right| \le B - A \quad \text{for all } x < y \text{ and all } u \text{ in } [a, b].$$

Proof: We apply the law of bounded change on $[x, y]$. Because the values of f' are in $[A, B]$, so is the difference quotient $(f(y) - f(x))/(y - x)$, which therefore cannot differ from any value of f' by more than $B - A$. ■

Corollary 7.6. *If f is uniformly differentiable on all sufficiently small subintervals of an interval I and if f' is uniformly continuous on I, then f is uniformly differentiable on I.*

Proof: For $\epsilon > 0$, the values of f' lie between $f'(x) - \epsilon$ and $f'(x) + \epsilon$ on each sufficiently small $[x, y]$ in I. Therefore, if $f(y) - f(x) = F(x, y)(y - x)$, Corollary 7.5 shows that $|F(x, y) - f'(x)| \leq 2\epsilon$ for $|y - x|$ sufficiently small. ■

The next consequence of the law of bounded change is needed for L'Hôpital's Rule. In it, A and B are constants, and f and g are uniformly differentiable on $[a, b]$.

Corollary 7.7. (*Generalized Law of Bounded Change*) *If $Ag' \leq f' \leq Bg'$ on $[a, b]$, then $A[g(b) - g(a)] \leq f(b) - f(a) \leq B[g(b) - g(a)]$.*

Proof: Apply the increasing function theorem to $Bg - f$ and $f - Ag$, and rearrange the resulting inequalities. ■

8. APPLICATION: L'HÔPITAL'S RULE. We now present a few examples in support of Tucker's contention that the increasing function theorem serves nicely to prove major theorems of the calculus that traditionally are derived from the mean value theorem [**6**]. We begin with L'Hôpital's Rule; see also [**2**]. There are two cases. In both, we assume that f and g are defined on a semi-infinite interval $[c, \infty)$ and are uniformly differentiable on each finite subinterval. We assume also that g and g' are positive.

Proposition 8.1. *If $f(x)$ and $g(x) \to 0$ and $f'(x)/g'(x) \to L$ as $x \to \infty$, then also $f(x)/g(x) \to L$ as $x \to \infty$.*

Proof: For $\epsilon > 0$, $L - \epsilon \leq f'/g' \leq L + \epsilon$ on $[p, \infty)$ if p is large enough. In that case, if $p \leq x \leq y$, the generalized law of bounded change ensures that

$$(L - \epsilon)(g(x) - g(y)) \leq f(x) - f(y) \leq (L + \epsilon)(g(x) - g(y)).$$

Because weak inequalities are preserved in the limit, if we let $y \to \infty$ and divide by $g(x) > 0$, we obtain $L - \epsilon \leq f(x)/g(x) \leq L + \epsilon$ for all $x \geq p$. ■

In the second case of L'Hôpital's Rule, it is common to assume also that $f(x) \to \infty$, but there is no need to do so.

Proposition 8.2. *If $g(x) \to \infty$ and $f'(x)/g'(x) \to L$ as $x \to \infty$, then also $f(x)/g(x) \to L$ as $x \to \infty$.*

Here too, the generalized law of bounded change is used only once. We note that if $f'(x)/g'(x)$ lies between $L - \epsilon/2$ and $L + \epsilon/2$ for $x \geq p$, then so does $(f(x) - f(p))/(g(x) - g(p))$. But to complete the argument, one has to be more artful than in the first case.

9. APPLICATION: DIFFERENTIATION UNDER THE INTEGRAL SIGN

Definition 9.1. A two-variable function f is *uniformly continuous* if for each $\epsilon > 0$, there is a $\delta > 0$ such that $|f(x', y') - f(x, y)| \leq \epsilon$ whenever both $|x' - x|$ and $|y' - y|$ are smaller than δ.

Remark 9.2. If the second coordinates in Definition 9.1 are equal, then the condition that guarantees that $|f(x', y) - f(x, y)| \leq \epsilon$ involves only the first coordinates: $|x' - x| \leq \delta$. That is, for each $\epsilon > 0$, one $\delta > 0$ works for all horizontal lines $y =$ constant. This simple observation is a key to our proof of Theorem 9.3.

Theorem 9.3. (*Differentiation Under the Integral Sign*) *Let f be defined on $Q = [a, b] \times [c, d]$ and uniformly continuous on $\{a\} \times [c, d]$. If f is uniformly differentiable on each $[a, b] \times \{y\}$ and its partial derivative f_x is uniformly continuous on Q, then f is uniformly continuous on Q and the integral of f_x over $[c, d]$ is a derivative for the integral of f over $[c, d]$.*

Proof: We assume the uniform continuity of f on Q. It can be be proved fairly easily using Corollary 7.3 but we prefer to focus here on the second part of the argument, which employs a less familiar application of the law of bounded change. Integrating $f(y, t) - f(x, t) = F(x, y, t)(y - x)$ over $[c, d]$, we see that to complete the proof, it suffices to demonstrate that the integral of $F(x, y, t) - f_x(x, t)$ over $[c, d]$ converges to 0 as $y - x \to 0$. To this end, it suffices to show that $|F(x, y, t) - f_x(x, t)|$ can be made less than any $\epsilon > 0$ by making $|y - x|$ less than some $\delta > 0$, independent of t.

It follows from Corollary 7.5 that $|F(x, y, t) - f_x(x, t)|$ is bounded for each t by any bound for $|f_x(v, t) - f_x(u, t)|$, for all u and v in $[x, y]$. By Remark 9.2, for any $\epsilon > 0$, there is a $\delta > 0$ such that if $|x - y| \leq \delta$, then, for all t in $[c, d]$, $|f_x(v, t) - f_x(u, t)| \leq \epsilon$ for all u and v in $[x, y]$. This is precisely what we need. ■

Proposition 9.3 also follows easily from reversal of order of integration and Corollary 7.3. Because reversal of order of integration is a simple consequence of the existence of the double integral of a uniformly continuous function, this provides a proof of Proposition 9.3 that uses the law of bounded change in a more familiar way.

10. HIGHER DIMENSIONS. In higher dimensions, there is no obvious counterpart to the increasing function theorem, and the mean value theorem is false even for a mapping from an interval to $\mathbb{R}^2$. Yet the law of bounded change generalizes almost without alteration if we regard convex sets as higher dimensional counterparts to intervals and read the proof as showing that if f is defined on an interval $[a, b]$, then any closed interval that contains $\{f'(u)(b - a) : u \in [a, b]\}$ also contains $f(b) - f(a)$.

Definition 10.1. A map f from a subset U of one normed linear space X to another Y is *uniformly differentiable* if there is a map Df from U to the set of bounded linear transformations from X to Y such that for each $\epsilon > 0$, $\|f(q) - f(p) - Df(p)(q - p)\|_Y \leq \epsilon \|q - p\|_X$ if $\|q - p\|_X$ is sufficiently small.

Proposition 10.2. *Using the notation in Definition* 10.1, *if U is convex then, for each p and q in U, $f(q) - f(p)$ belongs to every convex subset of Y that contains $Df(u)(q - p)$ for all u in U. Hence, if each $Df(u) : X \to Y$ is bounded by K on the unit sphere of X, then $\|f(q) - f(p)\|_Y \leq K\|q - p\|_X$.*

Proof: If p and q are in U, so is the line segment joining them. Hence, using a telescoping sum as in the proof of Theorem 7.1 and approximating each summand by the value of Df at a point on the segment, applied to $(q - p)/n$, we can approximate $f(q) - f(p)$ to arbitrary accuracy by an average of finitely many values of Df at points along the segment, applied to $q - p$. ■

11. AFTERWORD. We believe that this development, which is in the constructivist manner of Errett Bishop and L. E. J. Brouwer [**4**], produces proofs that are shorter and more transparent than those encountered in classical treatments. The idea of working with uniform rather than pointwise notions is a hallmark of the constructivist tradition.

For the one-dimensional case, our definition of differentiable function is a uniform version of a definition of Carathéodory. See [**3**] and the references therein. For a definition of this kind in higher dimensions, see [**1**].

REFERENCES

1. Acosta, E. and Delgado, C., Fréchet vs. Carathéodory, *Amer. Math. Monthly* **101** (1994) 332–338.
2. Boas, R.P., L'Hôpital's Rule without Mean Values, *Amer. Math. Monthly* **76** (1969) 1051–1053.
3. Kuhn, Stephen, The Derivative á la Carathéodory, *Amer. Math. Monthly* **98** (1991) 40–44.
4. Stolzenberg, Gabriel, review of *Foundations of Constructive Analysis* by Errett Bishop, *Bull. Amer. Math. Soc.* **76** (1970) 301–323.
5. Swann, Howard, Commentary on Rethinking Rigor in Calculus: The Role of the Mean Value Theorem, *Amer. Math. Monthly* **104** (1997) 241–245.
6. Tucker, Thomas, Rethinking Rigor in Calculus: The Role of the Mean Value Theorem, *Amer. Math. Monthly* **104** (1997) 231–240.

MARK BRIDGER (Columbia College '63) received his Ph.D. from Brandeis in 1967 as a student of Maurice Auslander. He is now working on constructive analysis, issues in the philosophy of science, and applications of technology to mathematics education. His other interests include music, bicycling, and gardening.
Northeastern University, Boston, MA 02115
bridger@neu.edu

GABRIEL STOLZENBERG received a Ph.D. from MIT in 1961. He worked first in several complex variables and then in constructive mathematics—his version of which is ordinary research done in a constructivist mind set with a minimum of attention to classical mathematics. His long experience with the classical/constructive gestalt switch is now being applied in a study of misreadings of humanists by scientists that will appear in an optimistically named volume, "After the Science Wars: Science and the Study of Science."
gabe@math.harvard.edu

The Answer is $2^n \cdot n!$ What's the Question?

Gary Gordon

1. INTRODUCTION. Allison, Cory, Heidi, and Zach are relaxing in the math coffee room. Allison says, "Did you see Jeopardy last night? I'm always excited when they have *Mathematics* as a category, but their math questions are always so lame!" Cory and Zach agree with Allison. Heidi says "Why don't they ever have an answer like $2^n n!$? Then the contestants would have to find a problem with $2^n n!$ as the answer. That might not be so easy, even for Alex."

Cory says "I've got it: What do you get when you multiply 2^n and $n!$?"

Ignoring Cory's correct but uninteresting solution, Allison continues, "Since we're mathematicians, our goal should be to find really good questions, using all sorts of math. We should each find our own question, one that uses our own field. Remember, we're trying to find good questions having $2^n n!$ as the answer."

The rest of this paper is devoted to the approaches taken by these four (fictitious, but aptly named) mathematicians. These approaches are from three different fields: algebra, geometry and combinatorics. This allows us to see the same problem from four different viewpoints, with each successive explanation encompassing the previous ones.

The algebraic and geometric approaches given here are well known, though perhaps not as widely known as they should be. The combinatorial approach is recent, incorporating a graph theory algorithm to produce the desired connections among these viewpoints.

In the hope of making this note more readable, the proofs are frequently sketched broadly. We hope enough details remain for the interested reader to fill in any missing steps (or look them up).

2. ALLISON'S TURN. Allison's specialty is algebra; she decides to find a group with $2^n n!$ elements. She needn't look far; the full symmetry group of an n-dimensional hypercube will do. There are n^2 mirror symmetry hyperplanes for an n-dimensional hypercube. One way to see this is to place the 2^n vertices of the hypercube at the points $(\pm 1, \ldots, \pm 1)$. Then the reflections of the hypercube occur in the hyperplanes whose equations are:

- $x_i = 0$ (the n coordinate hyperplanes),
- $x_i = x_j$ (there are $\binom{n}{2}$ of these), and
- $x_i = -x_j$ (there are $\binom{n}{2}$ of these, too),

where $1 \le i < j \le n$. See Figure 1 for a picture of the cube and its nine planes of reflection. In this figure, $x_i = 0$ (for $1 \le i \le 3$) corresponds to a plane that slices through the middle of four faces of the cube, while $x_i = x_j$ and $x_i = -x_j$ $(i \ne j)$ correspond to planes that contain the diagonals of two opposite faces and two edges of the cube.

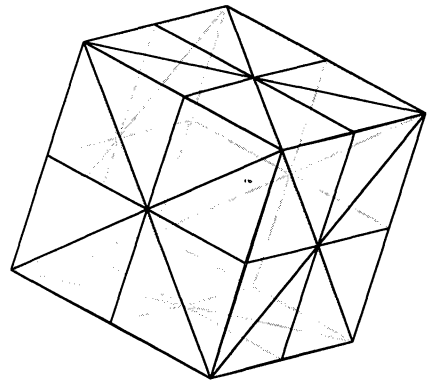


Figure 1. A cube and its planes of symmetry.

Let H_n denote the n-dimensional hypercube and let G_n be its symmetry group. Allison needs to show there are precisely $2^n n!$ elements in G_n. She decides to give an inductive argument. When $n = 1$, the hypercube is just a line segment, which has symmetry group $\mathbf{Z}_2$. In general, H_n has $2n$ *facets* (faces of dimension $n - 1$, each of which is the hypercube H_{n-1}). Allison constructs an arbitrary symmetry of H_n as follows:

1. Choose one of the $2n$ facets of H_n and move the hypercube so that this facet is on the bottom; the bottom of the hypercube is the facet that meets the x_n-axis at the point $(0, \ldots, 0, -1)$.
2. Use an element of G_{n-1} as a symmetry of this bottom face; in general, this symmetry permutes the other facets of H_n.

In 3 dimensions, this corresponds to picking one of the 6 faces of the cube for the bottom, then using one of the 8 symmetries of the square on this bottom face. This gives 48 symmetries, only 24 of which can be realized by rigid motions. The remaining symmetries are either reflections (there are 9, as we have already seen) or *rotary reflections*, i.e., reflections followed by rotations (there are evidently 15 of these). It is an interesting exercise to find explicit descriptions of these 15 rotary reflections for the cube.

Every symmetry of H_n can be obtained in this manner, since any symmetry must carry facets to facets. Then $|G_n| = 2n|G_{n-1}|$, which, together with the initial condition $|G_1| = 2$, gives $|G_n| = 2^n n!$

Allison's algebraic solution. Let G_n be the symmetry group of an n-dimensional hypercube. Then the order of G_n equals $2^n n!$

This approach gives us our answer in one way, but gives us very little information about the structure of the group G_n. We investigate the group in a bit more detail in the concluding section.

3. HEIDI'S TURN. Hyperplane arrangements form the background for the following classic problem, a favorite in mathematics contests:

- What is the largest number of regions produced when n lines are drawn in the plane?

This problem is also useful when introducing mathematical induction. The 3-dimensional version appeared as MONTHLY Problem E554 [8], where J. L. Woodbridge of Philadelphia asked:

- Show that n cuts can divide a cheese into as many as $(n + 1)(n^2 - n + 6)/6$ pieces.

Both of these questions are answered by a general formula discovered by L. Schläfli (published posthumously in 1901):

- The largest number of regions produced when n hyperplanes are drawn in d-dimensional space equals $\sum_{k=0}^{d} \binom{n}{k}$.

Counting the regions of various hyperplane arrangements is the beginning of a beautiful subject, with deep ties to algebra, topology, and combinatorics. A classic reference is [7]. Heidi has studied hyperplane arrangements and thinks she can use them to construct a good question. Since the algebraic approach has a strong geometric flavor, her first solution is to copy Allison's solution, without using groups. Recall that Allison used hyperplanes to understand the group G_n. If Heidi simply uses the same collection of hyperplanes Allison used (without even mentioning the hypercube), she will get a dissection of space into open n-dimensional regions. How many regions are produced by this hyperplane arrangement?

Thus, Heidi is concerned with counting the regions of the hyperplane arrangement whose equations are $x_i = 0$, $x_i = x_j$, and $x_i = -x_j$. Heidi needs to show that the number of regions is $2^n n!$.

What do these regions look like? Heidi finds it is easier to imagine the regions by intersecting them with a cube, as in Figure 1. Then a typical region is formed as follows: Let $O = (0, 0, 0)$ be the center of the cube, let $P_1 = (1, 0, 0)$ be the center of a face, let $P_2 = (1, 1, 0)$ be the center of an edge adjacent to this face, and finally let $P_3 = (1, 1, 1)$ be a vertex adjacent to this edge. Then the region is the tetrahedron whose vertices are the four points O, P_1, P_2, and P_3. There 6 choices for P_1, 4 choices for P_2, and 2 choices for P_3, giving us 48 regions.

In general, Heidi produces a region by picking the center of the hypercube, then picking the center of one of its $2n$ facets, then picking one of the $2(n - 1)$ faces (of dimension $n - 2$) of the chosen facet, and so on. At stage k, she chooses the center of one of the $2(n - k)$ facets surrounding an $n - k$–dimensional face of the hyperplane. The $n + 1$ points produced are the vertices of a simplex; this simplex is the intersection of a region of the hyperplane arrangement with the hypercube. See [2, §7.6] for more on this approach.

The number of regions of the arrangement is just $2n \cdot 2(n - 1) \cdots 4 \cdot 2 = 2^n n!$, so Heidi proudly announces

Heidi's hyperplane solution. Let A_n be the hyperplane arrangement given above. Then A_n decomposes space into $2^n n!$ regions of dimension n.

Heidi also wants to relate her approach to Allison's approach. She does so by exhibiting a one-to-one correspondence between the regions of the arrangement and the elements of the symmetry group G_n. To see this correspondence, first note that the symmetry group G_n acts on the regions of the arrangement. Now given any pair of regions A and B in her hyperplane arrangement, there is a unique element

$\sigma_{A,B} \in G_n$ with $\sigma_{A,B}(A) = B$ (G_n acts transitively on the arrangement). The existence of $\sigma_{A,B}$ follows because the regions themselves have no non-trivial symmetries–they are completely asymmetric. Then labeling an arbitrary region of the arrangement by the identity I of the group, Heidi gets her bijective correspondence between the regions $A_1, A_2, \ldots$ and the elements $\sigma_{I,A_1}, \sigma_{I,A_2}, \ldots$ Heidi has constructed a portion of the Cayley graph of G_n—the arrangement itself is equivalent to the geometric dual of the Cayley graph. See [**2**] for more on Cayley graphs.

4. ZACH'S TURN. If M is any $r \times k$ matrix, then the zonotope $Z(M)$ is a polytope in $\mathscr{R}^r$ given by

$$Z(M) = \left\{ \sum_{i=1}^{k} \lambda_i \mathbf{c}_i : -1 \leq \lambda_i \leq 1 \right\},$$

where $\mathbf{c}_i$ is the i^{th} column of M. Since the $\mathbf{c}_i$ are just an arbitrary collection of vectors in $\mathscr{R}^r$, the study of zonotopes is very natural in geometry or linear algebra. It is also fun to build 3-dimensional models of them using a good kit; Polydrons and Zometools are both good choices.

The hypercube H_n is a product of n intervals: $H_n = [-1,1] \times \cdots \times [-1,1]$. Zonotopes are projections of hypercubes; the definition shows how each interval $[-1,1]$ appears. The name *zonotope* arises from the fact that the facets determine 'zones' in space. Zonotopes are an important and well-studied class of polytopes with applications to oriented matroids, tiling problems and more. See [**1**], [**2**], and [**14**] for interesting connections among zonotopes, groups, geometry and oriented matroids.

Zach knows zonotopes; his idea is to find a class of zonotopes in which $2^n n!$ counts something. An obvious choice for a matrix whose columns generate the zonotope (given the first two solutions) is the matrix of normal vectors of the hyperplanes considered previously; $x_i = 0$, $x_i = x_j$, $x_i = -x_j$. For $n = 3$, a picture of the zonotope appears at the bottom of Figure 2. This polytope is a *rhombitruncated cuboctahedron*, one of the Archimedean solids. This zonotope arises as a truncation of a cube; the vertices of the cube correspond to the hexagons of the zonotope, the edges of the cube correspond to the squares and the faces (squares) of the cube correspond to the octagons. See Figure 2 for an illustration of how the truncation process produces the zonotope.

Zach constructs the $n \times n^2$ matrix M_n whose column vectors are normal to the hyperplanes (yes, the same hyperplanes Heidi used), and he calls the associated zonotope $Z(M_n)$. In general, order the columns of M_n as in the following example:

$$M_3 = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & -1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & -1 & 1 & -1 \end{bmatrix}$$

Zach now just counts the number of vertices of the zonotope $Z(M_n)$. Superimposing a picture of $Z(M_3)$ and the hyperplane arrangement A_3, Zach realizes that there is a one-to-one correspondence between the vertices of $Z(M_3)$ and the regions of the arrangement (see Figure 3). Note that each vertex of $Z(M_3)$ appears in the center of a region of the arrangement. Generalizing from 3-dimensions to n-dimensions (frequently dangerous but also frequently productive), he guesses

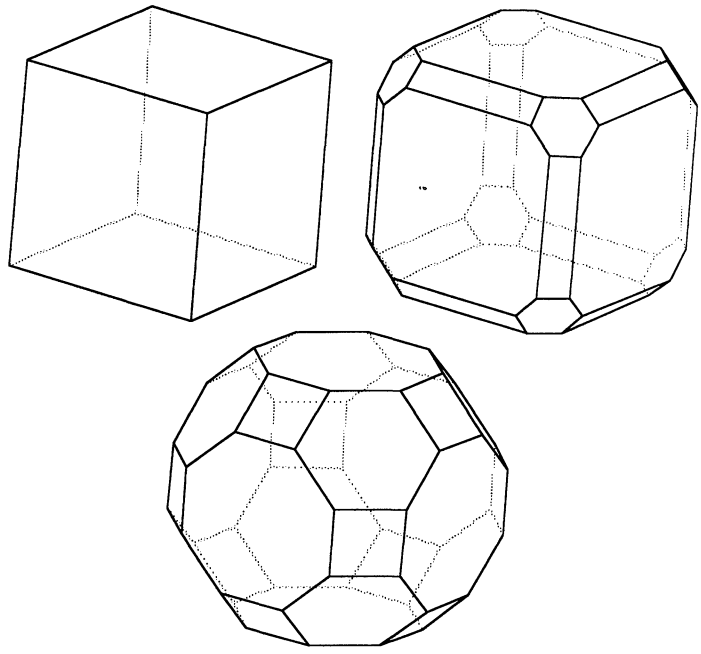


Figure 2. Truncating a cube to produce a zonotope.

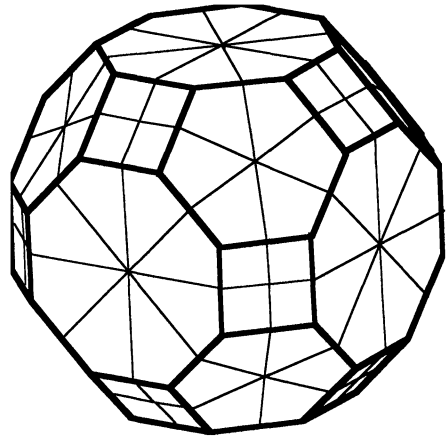


Figure 3. A zonotope and its planes of symmetry.

Zach's zonotope solution. The zonotope $Z(M_n)$ has $2^n n!$ vertices.

To see why this is true in general, Zach uses a little linear algebra. Here is a sketch of his ideas: First, he realizes that if the vector $\mathbf{v}$ in $\mathscr{R}^n$ is a vertex of the zonotope, then $\mathbf{v} = M_n\mathbf{p}$, where $\mathbf{p}$ is a column vector, each of whose n^2 entries equals 1 or -1. This follows from remembering that $Z(M_n) = \sum_{i=1}^{n^2} \lambda_i \mathbf{c}_i$, where $-1 \leq \lambda_i \leq 1$; a vertex of $Z(M_n)$ can ensue only when $\lambda_i = 1$ or -1. There are 2^{n^2} such ± 1 vectors $\mathbf{p}$, representing the vertices of an n^2-dimensional hypercube, the hypercube that is projected to $\cdot Z(M_n)$.

Each of these 2^{n^2} vertices of the hypercube produces a vector $\mathbf{v}$ that is now a possible vertex of $Z(M_n)$. Which $\mathbf{v}$ are the zonotope vertices? Zach figures out that any vector whose entries are some permutation of $\{1a_1, 3a_2, 5a_3, \ldots, (2n-1)a_n\}$, where each a_i equals 1 or -1, arises as $M_n\mathbf{p}$ for some vector of ± 1's.

Why do the vertices look like this? Every row of the matrix M_n has exactly $2n-1$ non-zero entries. Thus any coordinate of $M_n\mathbf{p}$ is bounded above by $2n-1$ and below by $-(2n-1)$. Once $\mathbf{p}$ has been chosen to make a certain coordinate of $M_n\mathbf{p}$ equal to $\pm(2n-1)$, the next largest (or smallest) coordinate value possible is $\pm(2n-3)$. The argument proceeds inductively; see [**3**] for a complete proof.

Conversely, each such $\mathbf{v}$ must be a vertex of the zonotope, since no such vector is a convex combination of other points in $Z(M_n)$. In particular, the matrix product $M_n\mathbf{p}$ can never produce any entry larger than $2n-1$ in absolute value. Thus, if $\pm(2n-1)$ appears as an entry in $\mathbf{v}$, then $\mathbf{v}$ could not be a non-trivial convex combination of other vertices of the zonotope. Finally, no other $\mathbf{v}$ is a vertex of $Z(M_n)$, since no other $\mathbf{v}$ contains $\pm(2n-1)$ as an entry.

The key for us is simply the following: since there are $n!$ permutations of $[1, 3, \ldots, (2n-1)]$ and 2^n possible ways to assign ± 1 to each entry, Zach has produced $2^n n!$, as required.

5. CORY'S TURN. Zach did his job, but he ducked an important question in considering the matrix products $M_n\mathbf{p}$:

Question 1. Of the 2^{n^2} possible sign vectors $\mathbf{p}$, which ones produce the $2^n n!$ vertices of the zonotope $Z(M_n)$?

The job of tying all the pieces together falls to Cory, our combinatorialist. Cory realizes that finding a question whose answer is $2^n n!$ is not much of a challenge; he could just find all permutations of $[\pm 1, \pm 3, \ldots, \pm(2n-1)]$ (as in Section 4) and be done. This involves combinatorics only superficially, however; he seeks a deeper connection, as do we.

The matrix M_n that Zach used reminds Cory of incidence matrices. Remember that every column of M_n has either exactly one non-zero entry (which equals 1) or exactly two non-zero entries (which either are both equal to 1 or have one entry equal to 1 and the other equal to -1). Cory decides to create a graph that has this matrix as its vertex-edge incidence matrix. Here's how the graph is defined: First, he labels n vertices of the graph he (modestly) calls C_n with the numbers $1, \ldots, n$ (corresponding to the rows of the matrix). Then the edges of the graph are filled in as follows:

1. Put a loop at every vertex (corresponding to the columns having one 1 and $n-1$ 0's),
2. Put an edge between every pair of vertices (corresponding to the columns having two 1's and $n-2$ 0's),
3. For each i and j such that $1 \le i < j \le n$, put a directed edge pointing from vertex i to vertex j (corresponding to the column having a 1 in position i, a -1 in position j and $n-2$ 0's).

Technically, C_n is a *mixed graph* because it mixes directed and undirected edges.

Cory has talked to Zach and knows about the connection between the matrix M_n and the zonotope $Z(M_n)$. In order to connect the graph C_n with the zonotope, he needs to interpret $M_n\mathbf{p}$, where $\mathbf{p} = [e_1, \ldots, e_{n^2}]^T$ and each $e_i = \pm 1$. He decides to use the entries of $\mathbf{p}$ to label some of the edges of C_n with ± 1. If column k of

M_n corresponds to an undirected edge in C_n, Cory labels this edge with e_k, which equals either 1 or -1. Similarly, if column k corresponds to a loop in C_n, he labels the loop with e_k. Finally, if column k of M_n corresponds to a directed edge from vertex i to vertex j, he leaves this directed edge unchanged if $e_k = 1$ and he reverses the direction if $e_k = -1$.

This process produces a *signed, oriented mixed graph* $C_n(\mathbf{p})$. See Figure 4 for an example with M_3 and the vector $\mathbf{p} = [1, -1, 1, -1, 1, 1, -1, 1, -1]^T$. To understand how this $\mathbf{p}$ produces the signs and arrows shown, recall the order of the

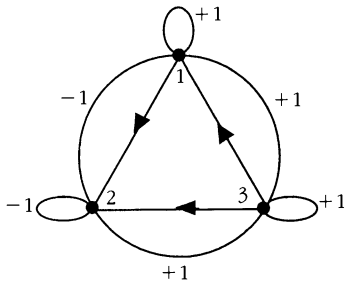


Figure 4. Signed mixed graph $G_3(\mathbf{p})$.

normal vectors appearing as the columns of M_3: The first three columns correspond to the loops at vertices 1, 2, and 3. Columns 4, 6, and 8 correspond to the (undirected) edges joining vertices 1 and 2, 1 and 3, and 2 and 3, respectively. Thus, for example, since $e_4 = -1$, the edge joining vertices 1 and 2 is signed with -1. Columns 5, 7, and 9 correspond to the directed edges joining vertices 1 and 2, 1 and 3, and 2 and 3, respectively. Thus, for example, since $e_7 = -1$ in the example, the edge directed from vertex 1 to vertex 3 is reversed.

There is a one-to-one correspondence between signed, oriented mixed graphs and the sign vectors $\mathbf{p}$. It is interesting to note that these sign vectors also arise naturally in the context of the hyperplane arrangements—Heidi's hyperplane approach to the problem. For any region R in the hyperplane arrangement, let v be a point in R. Thinking of v as a vector, compute the usual inner product of v with each of the n^2 normal vectors. Recording only whether these inner products are positive or negative produces a sign vector of length n^2; this simply records which side of each hyperplane the point v lies. Among all possible 2^{n^2} potential sign vectors that could arise in this way, it turns out that only $2^n n!$ do arise. These are precisely the sign vectors that produce vertices of the associated zonotope.

As anyone who has ever put up wallpaper knows, it is easy to move a bubble beneath the paper from place to place, but it's hard to get rid of the bubble. Cory has now shifted the problem of determining the $2^n n!$ vertices of the zonotope (or the regions of the hyperplane arrangement) to a graph theory problem.

Question 2. Of the 2^{n^2} possible oriented, signed mixed graphs $C_n(\mathbf{p})$, which ones produce the $2^n n!$ vertices of the zonotope $Z(M_n)$?

Cory's goal is to separate the two factors 2^n and $n!$ by having each one count a particular action. Now there are two aspects to $C_n(\mathbf{p})$: the orientation of the directed edges and the signing of the undirected edges. Cory decides to call a sign vector $\mathbf{p}$ *good* if it corresponds to a vertex of the zonotope. Cory notices a few things about these good $\mathbf{p}$: First, the orientation of $C_n(\mathbf{p})$ is *acyclic* when $\mathbf{p}$ is good.

This means that the vertices can be linearly ordered $v_1, v_2, \ldots, v_n$ so that every directed edge adjacent to v_1 points away from v_1, every directed edge adjacent to v_2 (except for the edge from v_1 to v_2) points away from v_2, and so on. In the language of tournaments, the directed edges of $C_n(\mathbf{p})$ form a *transitive* tournament —if a beats b and b beats c, then a beats c. There are $n!$ acyclic orientations—Cory is half-way home.

The second important point Cory realizes has to do with the signs on the undirected edges. When $\mathbf{p}$ is good, these signs can be obtained 'vertex by vertex' as follows: Assume $v_1, v_2, \ldots, v_n$ is the linear order and let $\mathbf{b} = [b_1, \ldots, b_n]^T$ be one of the 2^n vectors of 1's and -1's. If $b_1 = 1$, then he signs with a 1 all undirected edges incident to v_1 (including the loop); if $b_1 = -1$, then all undirected edges (including the loop) incident to v_n are signed with a -1. Now Cory removes v_1 or v_n from the list (depending on whether $b_1 = 1$ or -1), giving a new list of $n - 1$ vertices. He then repeats the process for b_2: if $b_2 = 1$, then he signs with a 1 all undirected edges incident to the first vertex on the new list (v_1 when $b_1 = -1$ and v_2 when $b_1 = 1$) (including the loop) *that have not previously been signed*; if $b_1 = -1$, then he signs with a -1 all undirected edges (including the loop) incident to the last vertex on the list (v_n when $b_1 = 1$ and v_{n-1} when $b_1 = -1$) *that have not previously been signed*. The process continues until all of the signs $b_1 - b_n$ have been processed. Since there are 2^n sign vectors $\mathbf{b}$ and each one gives a unique signing, Cory has the 2^n factor directly visible, too.

For example, if $\mathbf{b} = [1, -1, 1]^T$ and $v_1 = 3, v_2 = 1$, and $v_3 = 2$, he first paints the edges adjacent to vertex 3 with 1's, then paints the previously unsigned edges adjacent to 2 with -1's and finally paints the remaining unsigned edges adjacent to 1 with a 1; the only edge painted in this last step is the loop at vertex 1. See Figure 4.

Cory gives an inductive argument to show why this works; you can see it in [**3**]. The key for us is that this procedure for producing an oriented signed graph gives a direct link to our $2^n n!$ problem family: Since there are $n!$ orderings of the n vertices (to produce an acyclic orientation) and there are 2^n sign vectors $\mathbf{b}$ of length n, Cory immediately gets the answer of $2^n n!$.

Cory's combinatorial solution. There are precisely $2^n n!$ oriented, signed mixed graphs $C_n(\mathbf{p})$ corresponding to good sign vectors $\mathbf{p}$.

6. REBUTTAL. Allison, Heidi, Zach, and Cory meet in the coffee room to share their work. Allison sees a connection between the structure of the group G_n and the ubiquitous $2^n n!$

Allison's turn. Allison knows that the symmetry group G_n of a hypercube decomposes as a semi-direct product:

$$G_n = \mathbb{Z}_2^n \rtimes S_n.$$

She views the symmetry group of a hypercube as follows:

- First label the $2n$ facets of the hypercube by the symbols $1, 1^*, 2, 2^*, \ldots, n, n^*$, where the symbol i represents the facet contained in the hyperplane $x_i = 1$ and i^* represents the facet contained in the hyperplane $x_i = -1$.
- Choose a vertex v of the hypercube and record the ordered list of n facets incident to v. The starting point for the list can be determined uniquely by choosing the top or bottom facet first (whichever is incident to v) and fixing an orientation in space.

- Then an arbitrary symmetry of the hypercube can be broken down into two steps: First choose a permutation of the labels around the vertex v, then map v to some other vertex v'.

Allison explains to her colleagues: "The first step in this procedure can be accomplished by composing certain reflections through v (the collection of all reflections through v generates the symmetric group S_n), while the second step amounts to choosing an element from the normal subgroup $\mathbb{Z}_2^n$ to move from v to v' (accomplished by conjugating the permutation by the appropriate element of $\mathbb{Z}_2^n$, which is generated by the reflections perpendicular to the coordinate axes). Hey! Is anyone still awake?"

Cory's turn. Cory sees that the graph algorithm that produces the good sign vector **p** does (to a graph) more or less the same thing Allison just did (to the group G_n). To understand how the correspondence works, Cory reminds his colleagues that he first chooses an acyclic orientation of C_n, resulting in a permutation of the n vertices (which corresponds to the permutation of the n facets around the vertex v), then he picks a sign vector **b** of length n (which corresponds to mapping the vertex v in the hypercube to some other vertex). This gives a map of the $2^n n!$ elements of G_n to the collection of all $C_n(\mathbf{p})$ that are formed from good **p**'s.

Heidi adds, "It's really interesting how the four approaches used different areas of mathematics, but are so closely related. Let's send our solutions to Jeopardy!"

7. CONCLUDING REMARKS. The connections between hyperplane arrangements, zonotopes, and symmetry groups is explored in [**2**]. Further interpretations are examined in [**5**]. The sign vectors **p** considered here are maximal covectors in an associated oriented matroid [**1**]. The further connection with acyclic orientations of ordinary graphs is due to Curtis Greene [**4**], while the extension to mixed graphs appears in [**3**].

Signed graphs also provide a good way to understand the combinatorics of hyperplane arrangements. Tom Zaslavsky has developed a substantial theory for these combinatorial objects. See [**11**], [**12**], or [**13**] for a sample of this work. Many of his results generalize to other hyperplane arrangements. Zaslavsky has also considered hyperplane arrangements from a matroidal viewpoint. A very readable account appears in [**10**].

Many of the arguments given here can be rephrased using matrix groups; the reflections in the group G_n are easily represented by matrices. See [**2**] for more details on using linear algebra in this way.

ACKNOWLEDGMENT. Liz McMahon's comments substantially improved the exposition. Figures 1, 2, and 3 were created by the computer program KaleidoTile. The author also thanks Allison, Heidi, Cory, and Zach for their names.

REFERENCES

1. A. Björner, M. Las Vergnas, B. Sturmfels, N. White, G. Ziegler *Oriented Matroids*, Encyclopedia of Mathematics and Its Applications, Vol. 46, Cambridge Univ. Press, Cambridge, 1993.
2. H. S. M. Coxeter, *Regular Polytopes*, (Macmillan, 1963) third edition, Dover, New York, 1973.
3. G. Gordon, Hyperplane arrangements, hypercubes and mixed graphs, *Cong. Numer.* **126** (1997) 65–72.
4. C. Greene, Acyclic orientations, pp. 65–68 in *Higher Combinatorics* (M. Aigner ed.), Proc. NATO Advanced Study Institute (Berlin, 1976), D. Reidel, Boston, 1977.

5. C. Greene & T. Zaslavsky, On the interpretations of Whitney numbers through arrangements of hyperplanes, zonotopes, non-Radon partitions, and orientations of graphs, *Trans. Amer. Math. Soc.* **280** (1983) 97–126.
6. J. Humphreys, *Reflection groups and Coxeter groups*, Cambridge studies in advanced mathematics Vol. 29, Cambridge Univ. Press, Cambridge, 1992.
7. P. Orlik & H. Terao, *Arrangements of Hyperplanes*, Grundlehren der mathematischen Wissenschaften 300, Springer-Verlag, Berlin, 1992.
8. J. L. Woodbridge, Problem E 554, *Amer. Math. Monthly* **50** (1943) 59.
9. T. Zaslavsky, *Facing up to arrangements: Face-count formulas for partitions of space by hyperplanes*, Memoirs Amer. Math. Soc. **154**, 1975.
10. T. Zaslavsky, The geometry of root systems and signed graphs, *Amer. Math. Monthly* **88** (1981) 88–105.
11. T. Zaslavsky, Chromatic invariants of signed graphs, *Discrete Math.* **39** (1982) 215–228.
12. T. Zaslavsky, Signed graph coloring, *Discrete Math.* **42** (1982) 287–312.
13. T. Zaslavsky, Orientations of signed graphs, *Europ. J. Combinatorics* **12** (1991) 361–375.
14. G. Ziegler, *Lectures on polytopes*, Graduate Texts in Mathematics **152**, Springer-Verlag, New York, 1995.

GARY GORDON, a native Floridian and lifelong Dolphins fan, received his B.A. from the University of Florida in 1977 and his Ph.D. from the University of North Carolina in 1982. Like Allison, Cory, Heidi, and Zach, he is interested in combinatorics, geometry, and algebra. He loves watching baseball and playing softball. He also enjoys all sorts of games, but generally loses to his wife and frequent mathematical collaborator, Liz McMahon, and to his two daughters, Rebecca and Hannah.
Lafayette College, Easton, PA 18042
gordong@lafayette.edu

Conditional Convergence of Infinite Products

William F. Trench

In this article we revisit the classical subject of infinite products. For standard definitions and theorems on this subject see [**1**] or almost any textbook on complex analysis. We will restate parts of this material required to set the stage for our results, as follows.

The infinite product $P = \prod^{\infty}(1 + a_n)$ of complex numbers is said to *converge* if there is an integer N such that $1 + a_n \neq 0$ for $n \geq N$ and $\lim_{n\to\infty}\prod_{m=N}^{n}(1 + a_m)$ is finite and nonzero. This occurs if and only if the series $\sum_{m=N}^{\infty}\log(1 + a_m)$ converges.

We say that P *converges absolutely* if $\prod^{\infty}(1 + |a_n|)$ converges. If P converges absolutely then P converges, but the converse is false. The following theorem [**1**, p. 223] settles the question of absolute convergence of infinite products.

Theorem 1. *The infinite product* $\prod^{\infty}(1 + a_n)$ *converges absolutely if and only if* $\sum^{\infty}|a_n| < \infty$.

If P converges but $\prod^{\infty}(1 + |a_n|)$ does not, then we say that P *converges conditionally*. Conditional convergence of $\sum^{\infty}a_n$ does not imply conditional convergence of P. The following theorem [**1**, p. 225] seems to be the only general result along these lines, at least in the textbook literature.

Theorem 2. *If* $\sum^{\infty}|a_n|^2 < \infty$ *then* $\sum^{\infty}a_n$ *and* $\prod^{\infty}(1 + a_n)$ *converge or diverge together.*

Here we offer some other results concerning convergence of infinite products. Because of Theorem 1, these results are of interest only in the case where $\sum^{\infty}|a_n| = \infty$.

Theorem 3. *If there is a sequence* $\{r_n\}$ *such that*

$$\lim_{n\to\infty} r_n = 1 \tag{1}$$

and

$$\sum^{\infty}|r_n(1 + a_n) - r_{n+1}| < \infty, \tag{2}$$

then $\prod^{\infty}(1 + a_n)$ *converges.*

Proof: Let $g_n = r_n(1 + a_n) - r_{n+1}$. Then

$$\sum^{\infty}|g_n| < \infty \tag{3}$$

from (2), so $\lim_{n\to\infty}g_n = 0$ and therefore $\lim_{n\to\infty}a_n = 0$ by (1). Choose N so that r_n, $1 + a_n$ and $1 + g_n/r_{n+1}$ are nonzero if $n \geq N$. Now define $p_{N-1} = 1$ and

$$p_n = \prod_{m=N}^{n}(1 + a_m), \quad n \geq N.$$

If $n \geq N$ then $1 + a_n = p_n/p_{n-1}$, so $g_n = (r_n p_n/p_{n-1}) - r_{n+1}$, and therefore $p_n = r_{n+1} p_{n-1}(1 + g_n/r_{n+1})/r_n$, which implies that

$$p_n = \frac{r_{n+1}}{r_N} \prod_{m=N}^{n} (1 + g_m/r_{m+1}). \tag{4}$$

Since (1) and (3) imply that $\Sigma^{\infty}|g_m/r_{m+1}| < \infty$, Theorem 1 implies that the infinite product

$$Q = \prod_{m=N}^{\infty} (1 + g_m/r_{m+1})$$

converges; moreover $Q \neq 0$ because $1 + g_m/r_{m+1} \neq 0$ if $m \geq N$. Now (1) and (4) imply that $\lim_{n\to\infty} p_n = Q/r_N$ is finite and nonzero. ■

To apply this theorem we must exhibit a sequence $\{r_n\}$ that will enable us to obtain results even if $\Sigma^{\infty}|a_n| = \infty$. The following theorem provides a way to do this.

Theorem 4. *Suppose that for some positive integer q the sequences*

$$a_n^{(k)} = \sum_{m=n}^{\infty} a_m a_m^{(k-1)}, \quad k = 1, \ldots, q \ (\textit{with } a_m^{(0)} = 1),$$

are all defined, and

$$\sum \left| a_n a_n^{(q)} \right| < \infty. \tag{5}$$

Then $\prod^{\infty}(1 + a_n)$ *converges.*

Proof: Define

$$r_n^{(k)} = 1 + \sum_{j=1}^{k} (-1)^j a_n^{(j)}, \quad 1 \leq k \leq q.$$

We show by finite induction on k that

$$r_n^{(k)}(1 + a_n) - r_{n+1}^{(k)} = (-1)^k a_n a_n^{(k)} \tag{6}$$

for $1 \leq k \leq q$. Since $\lim_{n\to\infty} r_n^{(q)} = 1$ we can then set $k = q$ and conclude from (5) and Theorem 3 with $r_n = r_n^{(q)}$ that $\prod^{\infty}(1 + a_n)$ converges.

Since $r_n^{(1)} = 1 - a_n^{(1)}$ the left side of (6) with $k = 1$ is

$$\left(1 - a_n^{(1)}\right)(1 + a_n) - \left(1 - a_{n+1}^{(1)}\right) = a_n - a_n^{(1)} - a_n a_n^{(1)} + a_{n+1}^{(1)} = -a_n a_n^{(1)},$$

since $a_{n+1}^{(1)} + a_n = a_n^{(1)}$. This proves (6) for $k = 1$.

Now suppose that (6) holds if $1 \leq k < q - 1$. Since $r_n^{(k)} = r_n^{(k+1)} + (-1)^k a_n^{(k+1)}$, (6) implies that

$$\left(r_n^{(k+1)} + (-1)^k a_n^{(k+1)}\right)(1 + a_n) - r_{n+1}^{(k+1)} - (-1)^k a_{n+1}^{(k+1)} = (-1)^k a_n a_n^{(k)}.$$

Therefore

$$\begin{aligned} r_n^{(k+1)}(1 + a_n) - r_{n+1}^{(k+1)} &= (-1)^k \left(a_n a_n^{(k)} - a_n^{(k+1)} - a_n a_n^{(k+1)} + a_{n+1}^{(k+1)}\right) \\ &= (-1)^{(k+1)} a_n a_n^{(k+1)}, \end{aligned}$$

since $a_{n+1}^{(k+1)} + a_n a_n^{(k)} = a_n^{(k+1)}$. This completes the induction. ■

We now prepare for a specific application of Theorem 4. Henceforth Δ is the forward difference operator; thus, if $\{g_m\}$ is a sequence, then $\Delta g_m = g_{m+1} - g_m$,

while if G is a function of the continuous variable x then $\Delta G(x) = G(x+1) - G(x)$. Higher order forward differences are defined inductively; thus, if $\nu \geq 2$ is an integer, then

$$\Delta^{\nu} g_m = \Delta^{\nu-1} g_{m+1} - \Delta^{\nu-1} g_m = \sum_{r=0}^{\nu} (-1)^{r-\nu} \binom{\nu}{r} g_{m+r}.$$

A similar definition yields $\Delta^{\nu} G(x)$.

Lemma 1. *Suppose that t is a real number, not an integral multiple of 2π, and $\{g_m\}_{m=0}^{\infty}$ is a sequence such that $\lim_{m\to\infty} g_m = 0$ and*

$$\sum^{\infty} |\Delta^{\nu} g_m| < \infty \tag{7}$$

for some positive integer ν. Then $\sum^{\infty} g_m e^{imt}$ converges and

$$\sum_{m=0}^{\infty} g_m e^{imt} = (1 - e^{it})^{-\nu} \left[\sum_{s=0}^{\nu-1} A_s g_s + e^{i\nu t} \sum_{m=0}^{\infty} (\Delta^{\nu} g_m) e^{imt} \right], \tag{8}$$

where

$$A_s = \sum_{m=s}^{\nu-1} (-1)^{m-s} \binom{\nu}{m-s} e^{imt}, \quad 0 \leq s \leq \nu - 1. \tag{9}$$

Proof: Suppose that $M > 2\nu$ and let

$$S_M = (1 - e^{it})^{\nu} \sum_{m=0}^{M} g_m e^{imt}. \tag{10}$$

Since

$$(1 - e^{it})^{\nu} e^{imt} = \sum_{r=0}^{\nu} (-1)^r \binom{\nu}{r} e^{i(m+r)t},$$

we have

$$S_M = \sum_{m=0}^{M} g_m \sum_{r=0}^{\nu} (-1)^r \binom{\nu}{r} e^{i(m+r)t} = \sum_{r=0}^{\nu} (-1)^r \binom{\nu}{r} \sum_{m=0}^{M} g_m e^{i(m+r)t}$$
$$= \sum_{r=0}^{\nu} (-1)^r \binom{\nu}{r} \sum_{m=r}^{M+r} g_{m-r} e^{imt}.$$

Reversing the order of summation in the last sum yields

$$S_M = \sum_{m=0}^{\nu-1} \left(\sum_{r=0}^{m} (-1)^r \binom{\nu}{r} g_{m-r} \right) e^{imt} + \sum_{m=\nu}^{M} \left(\sum_{r=0}^{\nu} (-1)^r \binom{\nu}{r} g_{m-r} \right) e^{imt}$$
$$+ \sum_{m=M+1}^{M+\nu} \left(\sum_{r=m-M}^{\nu} (-1)^r \binom{\nu}{r} g_{m-r} \right) e^{imt}.$$

Since $\lim_{m\to\infty} g_m = 0$ the last sum on the right converges to 0 as $M \to \infty$. The second sum on the right is

$$\sum_{m=\nu}^{M} (\Delta^{\nu} g_{m-\nu}) e^{imt} = e^{i\nu t} \sum_{m=0}^{M-\nu} (\Delta^{\nu} g_m) e^{imt},$$

which converges as $M \to \infty$ because of (7). Therefore

$$\lim_{M\to\infty} S_M = S \equiv \sum_{m=0}^{\nu-1} \left(\sum_{r=0}^{m} (-1)^r \binom{\nu}{r} g_{m-r} \right) e^{imt} + e^{i\nu t} \sum_{m=0}^{\infty} (\Delta^{\nu} g_m) e^{imt},$$

which can also be written as

$$S = \sum_{s=0}^{\nu-1} A_s g_s + e^{i\nu t} \sum_{m=0}^{\infty} (\Delta^\nu g_m) e^{imt},$$

with A_s as in (9). This and (10) imply (8). ∎

Henceforth we write $G(x) = O(x^{-\alpha})$ to indicate that $x^\alpha G(x)$ remains bounded as $x \to \infty$.

Definition 1. Let $\mathcal{F}_\alpha$ be the set of infinitely differentiable functions F on $[1, \infty)$ such that

$$F^{(\nu)}(x) = O(x^{-\alpha-\nu}), \quad \nu = 0, 1, \ldots \tag{11}$$

For example, let $F(x) = u^\gamma(x)$, where u is a rational function with positive values on $[1, \infty)$ and a zero of order $p > 0$ at ∞; then F satisfies (11) with $\alpha = p\gamma$. To see this, we first recall that if $f = f(u)$ and $u = u(x)$, the formula of Faa di Bruno [2] for the derivatives of a composite function says that

$$\frac{d^\nu}{dx^\nu} f(u(x)) = \sum_{r=1}^{\nu} \frac{d^r}{du^r} f(u) \sum_r \frac{r!}{r_1! \cdots r_\nu!} \left(\frac{u'}{1!}\right)^{r_1} \left(\frac{u''}{2!}\right)^{r_2} \cdots \left(\frac{u^{(\nu)}}{\nu!}\right)^{r_\nu}, \tag{12}$$

where the prime denotes differentiation with respect to x. We are assuming here that the derivatives on the right of (12) exist. Here $u, \ldots, u^{(\nu)}$ are evaluated at x, and Σ_r is over all partitions of r as a sum of nonnegative integers,

$$r_1 + r_2 + \cdots + r_\nu = r, \tag{13}$$

such that

$$r_1 + 2r_2 + \cdots + \nu r_\nu = \nu. \tag{14}$$

Applying (12) with $f(u) = u^\gamma$ yields

$$F^{(\nu)}(x) = \sum_{r=1}^{\nu} (\gamma)^{(r)} u^{\gamma-r}(x) \sum_r \frac{r!}{r_1! \cdots r_\nu!} \left(\frac{u'(x)}{1!}\right)^{r_1} \left(\frac{u''(x)}{2!}\right)^{r_2} \cdots \left(\frac{u^{(\nu)(x)}}{\nu!}\right)^{r_\nu},$$

where $(\gamma)^{(r)} = \gamma(\gamma - 1) \cdots (\gamma - r + 1)$. Since $u^{(l)}(x) = O(x^{-p-l})$, it follows that

$$(u^{\gamma-r}(x))(u'(x))^{r_1}(u''(x))^{r_2} \cdots (u^{(\nu)}(x))^{r_\nu} = O(x^{-\lambda}),$$

where

$$\lambda = p(\gamma - r) + (p + 1)r_1 + (p + 2)r_2 + \cdots + (p + \nu)r_\nu = p\gamma + \nu$$

because of (13) and (14). This verifies (11) with $\alpha = p\gamma$.

For our purposes it is important to note that $\mathcal{F}_\alpha$ is a vector space over the complex numbers. Moreover, if $F_i \in \mathcal{F}_{\alpha_i}$, $i = 1, 2$, then $F_1 F_2 \in \mathcal{F}_{\alpha_1 + \alpha_2}$.

Lemma 2. *If $F \in \mathcal{F}_\alpha$ then*

$$\Delta^\nu F(x) = O(x^{-\alpha-\nu}), \quad \nu = 0, 1, 2, \ldots$$

Proof: We show that

$$|\Delta^\nu F(x)| \le K \max_{x < \xi < x + \nu} |F^{(\nu)}(\xi)|, \tag{15}$$

where K is a constant independent of F. Since $F^{(\nu)}(x) = O(x^{-\alpha-\nu})$ this implies the conclusion.

To verify (15), we note that if $x > 1$ and $r > 0$ then Taylor's theorem implies that

$$F(x+r) = \sum_{m=0}^{\nu-1} \frac{F^{(m)}(x)}{m!} r^m + \frac{F^{(\nu)}(\xi_r)}{\nu!} r^\nu,$$

where $x < \xi_r < x + r$. Since $\Delta^\nu F(x) = \sum_{r=0}^{\nu} (-1)^{r-\nu} \binom{\nu}{r} F(x+r)$, it follows that

$$\Delta^\nu F(x) = \sum_{m=0}^{\nu-1} \frac{F^{(m)}(x)}{m!} \left(\sum_{r=0}^{\nu} (-1)^{r-\nu} \binom{\nu}{r} r^m \right) + \frac{1}{\nu!} \sum_{r=0}^{\nu} (-1)^{r-\nu} \binom{\nu}{r} r^\nu F^{(\nu)}(\xi_r).$$

Since $\sum_{r=0}^{\nu} (-1)^{r-\nu} \binom{\nu}{r} r^m = 0$ for $m = 0, \ldots, \nu - 1$, we can now infer (15) with $K = (\sum_{r=0}^{\nu} \binom{\nu}{r} r^\nu)/\nu!$. ∎

Lemma 3. *Suppose that $F \in \mathcal{F}_\alpha$. Let ν be a fixed positive integer and let t be a real number, not an integral multiple of 2π. Then*

$$\sum_{m=n}^{\infty} F(m) e^{imt} = G(n) e^{int} + O(n^{-\alpha-\nu+1}),$$

where $G \in \mathcal{F}_\alpha$ (and G depends upon ν).

Proof: We write

$$\sum_{m=n}^{\infty} F(m) e^{imt} = e^{int} \sum_{m=0}^{\infty} F(n+m) e^{imt}. \tag{16}$$

From Lemma 2, $\Delta^\nu F(n+m) = O((n+m)^{-\alpha-\nu})$; that is, there is a constant A such that $|\Delta^\nu F(n+m)| < A(n+m)^{-\alpha-\nu}$ if $n + m > 0$. Therefore, if $n > 2$,

$$\sum_{m=0}^{\infty} |\Delta^\nu F(n+m)| < A \sum_{m=0}^{\infty} \frac{1}{(n+m)^\alpha} < A \sum_{m=0}^{\infty} \int_{n+m-1}^{n+m} \frac{dx}{(x+\alpha)^\nu}$$

$$= A \int_{n-1}^{\infty} \frac{dx}{(x+\alpha)^\nu} = O(n^{-\alpha-\nu+1}).$$

Applying Lemma 1 (specifically, (8)) with $g_m = F(n+m)$ and n fixed shows that

$$\sum_{m=0}^{\infty} F(n+m) e^{imt} = G(n) + O(n^{-\alpha-\nu+1})$$

with

$$G(x) = (1 - e^{it})^{-\nu} \sum_{s=0}^{\nu-1} A_s F(x+s),$$

so $G \in \mathcal{F}_\alpha$. Now (16) implies the conclusion. ∎

The following theorem shows that Theorem 4 has nontrivial applications for every positive integer q.

Theorem 5. *Suppose that*

$$a_n = f(n) e^{in\theta}, \quad n = 1, 2, 3, \ldots, \tag{17}$$

where $f \in \mathcal{F}_\gamma$ for some $\gamma \in (0, 1]$, and let q be the smallest integer such that

$$(q+1)\gamma > 1. \tag{18}$$

Then the infinite product $P = \prod^{\infty}(1 + a_n)$ converges if θ is not of the form $2k\pi/r$ with k an integer and $r \in \{1, \ldots, q\}$.

Proof: We show by finite induction on p that if $p = 1, \dots, q$ then

$$a_n a_n^{(p)} = f_p(n) e^{i(p+1)n\theta} + O(n^{-(p+1)\gamma - q + p}) \tag{19}$$

where $f_p \in \mathscr{F}_{(p+1)\gamma}$. In particular, (19) with $p = q$ implies that $a_n a_n^{(q)} = O(n^{-(q+1)\gamma})$, so (18) implies (5) and P converges, by Theorem 4.

From (17) and Lemma 3 with $t = \theta$, $F = f$, $\alpha = \gamma$, and $\nu = q$,

$$a_n^{(1)} = \sum_{m=n}^{\infty} f(m) e^{im\theta} = G_1(n) e^{in\theta} + O(n^{-\gamma - q + 1}),$$

with $G_1 \in \mathscr{F}_\gamma$. Therefore $a_n a_n^{(1)} = f(n) e^{in\theta}(G_1(n) e^{in\theta} + O(n^{-\gamma - q + 1}))$. Since $f \in \mathscr{F}_\gamma$, this can be rewritten as $a_n a_n^{(1)} = f_1(n) e^{2in\theta} + O(n^{-2\gamma - q + 1})$, with $f_1 = f G_1 \in \mathscr{F}_{2\gamma}$. This establishes (19) with $p = 1$, so we are finished if $q = 1$.

Now suppose that $q > 1$ and (19) holds if $1 \le p < q$. Since $(p+1)\theta$ is by assumption not an integral multiple of 2π, Lemma 3 with $t = (p+1)\theta$, $F = f_p$, $\alpha = (p+1)\gamma$, and $\nu = q - p$ implies that

$$\sum_{m=n}^{\infty} f_p(m) e^{i(p+1)m\theta} = G_p(n) e^{i(p+1)n\theta} + O(n^{-(p+1)\gamma - q + p + 1}),$$

where $G_p \in \mathscr{F}_{(p+1)\gamma}$. This and (19) imply that

$$a_n^{(p+1)} \equiv \sum_{m=n}^{\infty} a_m a_m^{(p)} = G_p(n) e^{i(p+1)n\theta} + O(n^{-(p+1)\gamma - q + p + 1}),$$

so

$$a_n a_n^{(p+1)} = f(n) e^{in\theta}\left(G_p(n) e^{i(p+1)n\theta} + O(n^{-(p+1)\gamma - q + p + 1})\right).$$

Since $f \in \mathscr{F}_\gamma$, this can be rewritten as

$$a_n a_n^{(p+1)} = f_{p+1}(n) e^{i(p+2)n\theta} + O(n^{-(p+2)\gamma - q + p + 1}),$$

with $f_{p+1} = f G_p \in \mathscr{F}_{(p+2)\gamma}$. This completes the induction. ∎

Corollary 1. *Suppose that $\{a_n\}^\infty$ is as defined in Theorem 5. Then the infinite product $\prod^\infty(1 + a_n)$ converges if θ is not a rational multiple of 2π.*

Corollary 2. *Suppose that $\alpha > 0$ and R is a rational function such that $R(x) > 0$ on $[N, \infty)$ (N = integer) and $\lim_{n\to\infty} R(x) = 0$. Then the infinite product $\prod_{n=N}^\infty (1 + (R(n))^\alpha e^{in\theta})$ converges if θ is not a rational multiple of 2π.*

Corollary 3. *The infinite product $\prod^\infty(1 + n^{-\alpha} e^{in\theta})$ converges if $\alpha > 0$ and θ is not a rational multiple of 2π.*

REFERENCES

1. K. Knopp, *Theory and Application of Infinite Series*, Hafner Publishing Company, New York, 1947.
2. Ch.-J. de La Vallee Poussin, *Cours d'analyse infinitesimale*, Vol. 1, 12th Ed., Libraire Universitaire Louvain, Gauthier-Villars, Paris, 1959.

WILLIAM F. TRENCH is retired and lives with his wife Beverly and their dog Zachary in a forest in Divide, Colorado (elevation 9200 feet). He received his Ph.D in 1958 from the University of Pennsylvania. After working in nonacademic positions from 1953 to 1964, he taught at Drexel University from 1964 to 1986. From 1986 to 1997 he was Andrew G. Cowles Distinguished Professor at Trinity University, San Antonio, Texas. His mathematical interests include differential and difference equations, linear algebra, special functions, and classical analysis.

413 Lake Drive West, Divide, CO 80814

wtrench@trinity.edu

H. J. S. Smith and the Fermat Two Squares Theorem

F. W. Clarke, W. N. Everitt, L. L. Littlejohn, and S. J. R. Vorster

This article is dedicated to Professor P. R. Halmos

1. INTRODUCTION. In his remarkable book *A Mathematician's Apology*, G. H. Hardy wrote [**12**, p. 97]

> Another famous and beautiful theorem is Fermat's 'two square' theorem. The primes may (if we ignore the special prime 2) be arranged in two classes; the primes
>
> $$5, 13, 17, 29, 37, 41, \ldots$$
>
> which leave remainder 1 when divided by 4, and the primes
>
> $$3, 7, 11, 19, 23, 31, \ldots$$
>
> which leave remainder 3. All the primes of the first class, and none of the second, can be expressed as the sum of two squares: thus
>
> $$5 = 1^2 + 2^2, \quad 13 = 2^2 + 3^2$$
> $$17 = 1^2 + 4^2, \quad 29 = 2^2 + 5^2$$
>
> but 3, 7, 11 and 19 are not expressible in this way (as the reader may check by trial). This is Fermat's theorem, which is ranked, very justly, as one of the finest of arithmetic. Unfortunately there is no proof within the comprehension of anybody but a fairly expert mathematician.

The history of this theorem of Fermat is given in detail by Dickson [**7**, 224–237]. Dickson names the theorem after Girard, who discussed the result in 1632; however the common practice now is to attribute the result to Fermat, who stated in 1659 that he possessed an irrefutable proof by the method of infinite descent; see [**7**, p. 228] and [**2**, p. 89]. The first recorded proof is due to Euler given in 1749 [**7**, pp. 230–231]; Bell writes, "It was first proved by the great Euler in 1749 after he had struggled, off and on, for *seven years* to find a proof" [**2**, p.89]. The first proof that such prime numbers can be *uniquely* represented as the sum of squares of two positive integers was given by Gauss in 1801 [**7**, p. 233]. See also the account of the two squares theorem of Fermat in the books by Burton [**4**, Chapter 12, Section 2], and Hardy and Wright [**14**, Chapter XX].

The last sentence in the quotation from Hardy is significant. Hardy had an interest in the classification of proof; see, in particular, [**13**, p. 6] in connection with the "elementary" proof of inequalities. In this context the word *elementary* must not be confused with the words *obvious* or *easy*; many of the elementary proofs in [**13**] are subtle, ingenious, and far from obvious. When Hardy wrote [**12**] he was, more than likely, not aware that an elementary proof of this theorem of Fermat had been given in 1855 by H. J. S. Smith, one of his predecessors in the Savilian Chair of Geometry in the University of Oxford. This simple but remarkable proof of Smith is within the comprehension of those with knowledge of elementary

algebra, including simple properties of determinants, and the fundamental theorem of arithmetic [**6**, Chapter I, Section 4]. The proof is also remarkable for giving a construction that permits one to compute the integers of the two squares representation.

In this paper we give Smith's proof of the theorem of Fermat and present what is, possibly, a new elementary proof of the uniqueness of the two squares representation, but now using Smith's ideas and method. This uniqueness proof involves the Euler Criterion [**8**, Section 11] for solutions of the quadratic equation $x^2 \equiv -1 \pmod{p}$; we present a new existence proof that leads to a constructable solution of this equation.

The original paper of Smith [**20**] is (the good news) only 2 pages long but is (the bad news for most of us) written in Latin; see also the collected works of Smith [**21**], in which [**20**] appears as the second contribution. The Smith proof has not gone entirely without notice; Chrystal [**5**, p. 471] reproduces the proof in English, as does, in part, Dickson [**7**, pp. 240–241]; Davenport mentions the proof [**6**, p. 122] but does not give complete details. Barnes [**1**] gives an exposition of Smith's existence theorem, and establishes the connection between the Smith palindromic continuant and the Euler Criterion (see our Theorems 1 and 2 and their proofs).

Both Serret [**19**] and Hermite [**17**] use ideas similar to the Smith method [**20**] to give an algorithm for finding the integers in the two squares representation of the theorem of Fermat. This method was subsequently improved by Brillhart [**3**] to give an impressively fast numerical procedure to determine the representation; as an example the Brillhart method gives

$$10^{50} + 577 = 761106534380835424545040 1^2 + 648626890687392164224542 4^2. \tag{1.1}$$

The two squares theorem of Fermat continues to attract attention; see the recent contributions by Ewell [**9**], Heath-Brown [**16**], Wagon [**22**], and Zagier [**23**].

In Section 2 we give formal statements of the results to be proved by the Smith methods. In Section 3 we give a brief account of the life of Henry Smith. In Section 4 there is a definition and statement of the properties of continuants. The remaining sections are devoted to proofs of the results. Lastly, in an appendix, we reproduce the original Smith paper [**20**].

At the end of Sections 6, 7, and 8 we exemplify the general results by considering the case $p = 13$, and other cases including the example in (1.1).

2. STATEMENT OF RESULTS. Let $\mathbb{N} := \{1, 2, 3, \ldots\}$ and $\mathbb{P} := \{p \in \mathbb{N} : p \text{ is a prime number}\}$.

Theorem 1 [Fermat and Gauss]. *Let $p \in \mathbb{P}$ with $p \equiv 1 \pmod{4}$. Then there exist two unique, positive, co-prime integers $u, v \in \mathbb{N}$ such that*

$$p = u^2 + v^2.$$

Proof: See Sections 6 and 7. ∎

Theorem 2 [The Euler Criterion]. *Let $p \in \mathbb{P}$ with $p \equiv 1 \pmod{4}$. Then*

1. *The quadratic equation*

$$x^2 \equiv -1 \pmod{p} \tag{2.1}$$

has two unique solutions $x_0, x_1 \in \mathbb{N}$ such that

$$1 < x_0 < (p-1)/2 \quad \text{and} \quad (p-1)/2 < x_1 < p,$$

with $x_1 = p - x_0$.

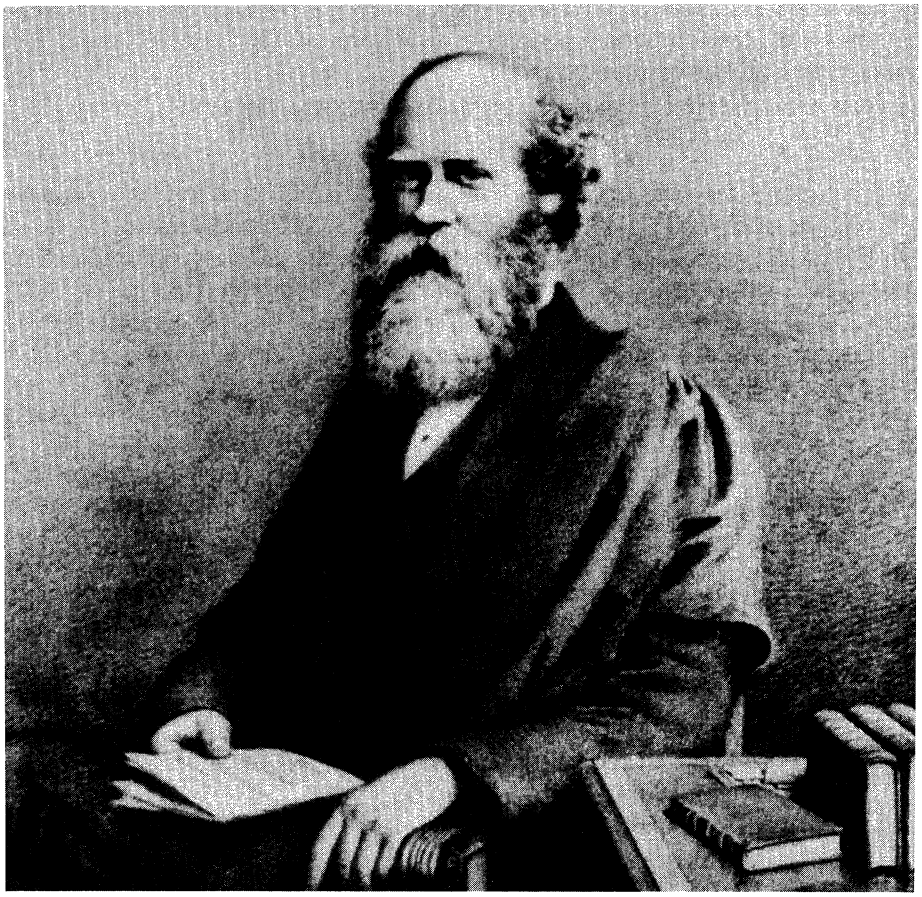
2. *All other solutions of (2.1) are congruent to x_0 or $x_1 \pmod{p}$.*

Proof: See Section 8. ■

For a detailed discussion of the Euler Criterion see the book by Dudley [**8**, pp. 85-86].

2. HENRY JOHN STEPHEN SMITH. Henry Smith was born on 2 November 1826 in Dublin, Ireland. His father died soon afterwards and the widow moved with her family to England. Smith was educated first by his mother and then by a succession of private tutors, before spending three years at Rugby School; from this School he gained entry to the University of Oxford, in 1844, by winning the top scholarship to Balliol College. In 1848 at Oxford he gained first class honours in both classics and mathematics; he also won the major University prizes in both these subjects, the Ireland scholarship in classics, and the Senior Mathematical Scholarship in mathematics.

In 1849 the Balliol College fellowships in classics and mathematics fell vacant; until this time Smith seems to have been undecided as to whether to follow a career in classics or mathematics, but seems to have settled at this time on mathematics. His first paper, on geometry, dates from the next year.



Henry John Stephen Smith (1826–1883).

Smith completed his first paper on number theory in 1854 and published it the following year in *Crelle's Journal* [**20**]. Unusually, even for that time, it was written in Latin, perhaps in homage to Carl Friedrich Gauss, whose *Disquisitiones arithmeticae* served as an inspiration.

In 1859 Smith was elected to Fellowship of the Royal Society of London and then, in 1860, to the Savilian Chair of Geometry in the University of Oxford; two of his eventual successors to this Chair were G. H. Hardy and E. C. Titchmarsh.

Henry Smith died in Oxford in 1883.

It is puzzling that Henry Smith's work and name are so little known, even amongst those who make regular use of the ideas that he introduced; this point is discussed by Keith Hannabuss in his paper *The mathematician the world forgot* [**10**]. Several historians of mathematics have ranked Smith with Cayley and Sylvester among the great pure mathematicians of the nineteenth century. In 1861 Smith proved the existence and uniqueness of what is now called the *Smith Normal Form* of a matrix with integer entries. This result has subsequently been used to prove the cyclic decomposition for modules, but Smith's first application was to determine when linear Diophantine equations admit solutions, settling a longstanding problem first studied by Greek mathematicians. His remarkable contributions to, and his panoramic knowledge of, the theory of numbers can be seen in the monumental *Report on the theory of numbers*, reproduced in [**21**]. In this area, in 1868, he shared the Steiner Prize of the Royal Academy of Sciences in Berlin for his solution of a geometric problem that involved the representation of integers as a sum of squares.

Not so well known is Smith's early contribution to measure theory and integration in his paper of 1875 *On the integration of discontinuous functions*; see [**21**, paper 25]. There, Smith introduced the first example of what is now called a Cantor set; Cantor's own example appeared eight years later and was not presented as his own discovery. Smith's example divides an interval into $m > 2$ subintervals, and then keeps repeating this process to each remaining subinterval, except the last. Smith also seems to have been the first mathematician to perceive the connection between measure and integral. However, his paper received less attention than it deserved, owing to an inaccurate review in the *Fortschritte der Mathematik*. In his history of integration, see [**15**, pp. 37, 40], Thomas Hawkins has remarked:

> Probably the development of a measure-theoretic viewpoint within integration theory would have been accelerated had the contents of Smith's paper been known to mathematicians whose interest in the theory was less tangential than Smith's.

For an informed discussion on the contents of this paper of Smith, and for the development of the ideas therein to higher dimensions, see [**10**] and, especially, [**11**].

4. CONTINUANTS. Continuants are closely connected with continued fractions, as noted by Smith at the beginning of [**20**]. There is a detailed and elegant account of this connection in Chrystal [**5**, Chapter XXXIV, Sections 4–11]. However Smith uses only continuants in his paper and uses determinants to define them; for this definition see [**5**, Chapter XXXIV, Section 11] and the reference therein to the

remarkable history of determinants by Muir and Metzler [**18**, Chapters III and XIII]. We follow Smith and make the

Definition 1. For $n \in \mathbb{N}$ let $q_r \in \mathbb{N}$ $(r = 1, 2, \ldots, n)$; then define $[\cdot]: \mathbb{N}^n \to \mathbb{N}$ by the determinant

$$[q_1, q_2, q_3, \ldots, q_{n-1}, q_n] := \begin{vmatrix} q_1 & 1 & 0 & \cdots & 0 & 0 \\ -1 & q_2 & 1 & \cdots & 0 & 0 \\ 0 & -1 & q_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & q_{n-1} & 1 \\ 0 & 0 & 0 & \cdots & -1 & q_n \end{vmatrix} \tag{4.1}$$

We note that

$$[q_1] = q_1, \quad [q_1, q_2] = q_1 q_2 + 1, \quad \text{and } [q_1, q_2, q_3] = q_1 q_2 q_3 + q_1 + q_3. \tag{4.2}$$

Lemma 1. *Let $n \in \mathbb{N}$ with $n \geq 2$. Then*

1. $[q_1, q_2, \ldots, q_n] = [q_1][q_2, q_3, \ldots, q_n] + [q_3, \ldots, q_n]$
2. $[q_1, q_2, \ldots, q_n] \in \mathbb{N}$
3. $[q_1, q_2, \ldots, q_n] = [q_n, \ldots, q_2, q_1]$
4. $[q_2, q_3, \ldots, q_n] < [q_1, q_2, \ldots, q_n]$
5. $[q_2, q_3, \ldots, q_n]$ *and* $[q_1, q_2, \ldots, q_n]$ *are co-prime integers*
6. $$[q_1, \ldots, q_{s-1}, q_s, q_{s+1}, q_{s+2}, \ldots, q_n] = [q_1, \ldots, q_{s-1}, q_s][q_{s+1}, q_{s+2}, \ldots, q_n] + [q_1, \ldots, q_{s-1}][q_{s+2}, \ldots, q_n].$$

Proof: Note that if in any formula in Lemma 1 an empty continuant appears then it is convenient, and consistent, to give such a continuant the value 1.

1. Expand the determinant (4.1) by the first row.
2. Use (4.2), property 1 and mathematical induction.
3. Standard property of determinants.
4. Use properties 1 and 2.
5. Use property 1.
6. Use the Laplace expansion on (4.1) centred on row s; see also the proof in [**5**, Chapter XXXIV, Section 6].

5. THE EUCLIDEAN ALGORITHM

Algorithm 1. *Let $r, s \in \mathbb{N}$ be co-prime with $s < r$ and write*

$$\frac{r}{s} = q_1 + \frac{t}{s} \quad (0 < t < s), \frac{s}{t} = q_2 + \frac{u}{t} \quad (0 \leq u < t), \ldots, \frac{v}{w} = q_n + \frac{0}{w} = q_n \tag{5.1}$$

for some $n \in \mathbb{N}$ with $n \geq 2$, $q_i \in \mathbb{N}$ $(i = 1, 2, \ldots, n)$, and $q_n \geq 2$.

Thus a rational number $r/s > 1$ is associated with a set of positive integers $\{q_1, q_2, \ldots, q_n\}$ satisfying the properties in (5.1). Conversely we have

Lemma 2. *Let a set of positive integers* $\{q_1, q_2, \ldots, q_n\}$ *be given with* $n \geq 2$ *and* $q_n \geq 2$. *Then there is a unique rational number* $r/s > 1$ *whose Euclidean algorithm yields the set* $\{q_1, q_2, \ldots, q_n\}$; *moreover,* r/s *is determined by*

$$\frac{r}{s} = \frac{[q_1, q_2, \ldots, q_n]}{[q_2, q_3, \ldots, q_n]}. \tag{5.2}$$

Here r *and* s *are co-prime and given by*

$$r = [q_1, q_2, \ldots, q_n] \quad \textit{and} \quad s = [q_2, q_3, \ldots, q_n]. \tag{5.3}$$

Proof: Define r/s by (5.2) and apply property 1 of Lemma 1 n times. The result (5.3) follows from property 5 of Lemma 1. ∎

6. THE SMITH PROOF OF THE FERMAT THEOREM

Proof: Let $p \in \mathbb{P}$ with $p \equiv 1 \pmod 4$ and write $p = 4r + 1$. Let the number μ be taken arbitrarily from the set of positive integers $\{1, 2, \ldots, 2r\}$ and consider the corresponding set of rational numbers $\{p/\mu\}$, noting that $2 < p/\mu \leq p$. If we apply Algorithm 1 to p/μ we obtain a representation of the form

$$\frac{p}{\mu} = \frac{[q_1, q_2, \ldots, q_n]}{[q_2, \ldots, q_n]}. \tag{6.1}$$

Of course, the integer n and the set $\{q_1, q_2, \ldots, q_n\}$ depend upon the particular choice of μ. From property 5 of Lemma 1 we obtain

$$p = [q_1, q_2, \ldots, q_n] \quad \text{and} \quad \mu = [q_2, \ldots, q_n]. \tag{6.2}$$

From Algorithm 1 and from property 1 of Lemma 1, since $p/\mu > 2$, it follows that, in the representation (6.2),

$$q_1 \geq 2 \quad \text{and} \quad q_n \geq 2. \tag{6.3}$$

Now take one of the rational numbers p/μ with

$$\mu \in \{2, 3, \ldots, 2r\};$$

then we have the following chain of argument, using property 3 of Lemma 1 and (6.1),

$$\frac{p}{\mu} = \frac{[q_1, q_2, \ldots, q_n]}{[q_2, \ldots, q_n]} \Rightarrow [q_1, q_2, \ldots, q_n] = p = [q_n, q_{n-1}, \ldots, q_1]$$

$$\Rightarrow \frac{[q_n, q_{n-1}, \ldots, q_1]}{[q_{n-1}, \ldots, q_1]} = \frac{p}{\nu}, \tag{6.4}$$

(say). It follows from (6.3), Lemma 2, and property 1 of Lemma 1 that $1 < \nu < p/2$, so $\nu \in \{2, 3, \ldots, 2r\}$. Thus the chain of argument that gave (6.4) can be reversed, starting with ν and finishing with μ.

This argument pairs off the elements of the set $\{2, 3, \ldots, 2r\}$ and gives each member μ of the set a unique mate ν in the set. However this set contains an odd number of elements so there must exist at least one member, say λ, that mates with itself in the chain (6.4). For this λ we obtain from (6.4)

$$\frac{[q_1, q_2, \ldots, q_n]}{[q_2, \ldots, q_n]} = \frac{p}{\lambda} = \frac{[q_n, q_{n-1}, \ldots, q_1]}{[q_{n-1}, \ldots, q_1]}. \tag{6.5}$$

Now apply Algorithm 1 to both sides of (6.5) to give a representation

$$p = [q_1, q_2, \ldots, q_n], \tag{6.6}$$

with the palindromic property, and with (6.3) holding,

$$q_i = q_{n+1-i} \quad (i = 1, 2, \ldots, n). \tag{6.7}$$

If, in (6.7), $n = 2t + 1$ is odd then $n \geq 3$ and the representation (6.6) takes the form, for $s \geq 2$,

$$p = [q_1, \ldots, q_{s-1}, q_s, q_{s-1}, \ldots, q_1].$$

Now apply property 6 of Lemma 1 to give

$$p = [q_1, \ldots, q_{s-1}, q_s][q_{s-1}, \ldots, q_1] + [q_1, \ldots, q_{s-1}][q_{s-2}, \ldots, q_1].$$

Other properties of Lemma 1 permit us to write

$$p = [q_1, \ldots, q_{s-1}]\{[q_1, \ldots, q_{s-1}, q_s] + [q_{s-2}, \ldots, q_1]\},$$

which represents the prime number p as the product of two factors that, using (6.3), are both greater than 1; this is a contradiction to $p \in \mathbb{P}$.

Thus in (6.7) the integer $n = 2t$ must be even and so (6.6) takes the form, for $s \geq 1$,

$$p = [q_1, \ldots, q_s, q_s, \ldots, q_1]$$

with $q_1 \geq 2$ from (6.3). Now apply property 6 of Lemma 1 to give

$$p = [q_1, \ldots, q_s][q_s, \ldots, q_1] + [q_1, \ldots, q_{s-1}][q_{s-1}, \ldots, q_1]$$

and then

$$p = [q_1, \ldots, q_s]^2 + [q_1, \ldots, q_{s-1}]^2.$$

From property 1 it follows that $[q_1, \ldots, q_{s-1}]$ and $[q_1, \ldots, q_s]$ are co-prime.

This completes Smith's proof of the Fermat part of Theorem 1. ■

Consider the case $p = 13$. Then $\mu \in \{2, 3, 4, 5, 6\}$ and the application of the Euclidean algorithm to each choice of μ gives

$$\begin{array}{llllll}
\mu = 2 & n = 2 & q_1 = 6 & q_2 = 2 & & \\
\mu = 3 & n = 2 & q_1 = 4 & q_2 = 3 & & \\
\mu = 4 & n = 2 & q_1 = 3 & q_2 = 4 & & \\
\mu = 5 & n = 4 & q_1 = 2 & q_2 = 1 & q_3 = 1 & q_4 = 2 \\
\mu = 6 & n = 2 & q_1 = 2 & q_2 = 6. & &
\end{array}$$

Thus in the Smith pairing, 2 pairs with 6, 3 pairs with 4, and 5 pairs with itself. The palindromic continuant given by 5 then yields the two squares result

$$13 = [2, 1, 1, 2] = [2, 1][1, 2] + [2][2] = [2, 1]^2 + [2]^2 = 3^2 + 2^2.$$

7. A COROLLARY. We have

Corollary 1. *Let $p \in \mathbb{P}$ with $p = 4r + 1$. Then there are exactly $2r$ distinct continuant representations of p*

$$p = [q_1, \ldots, q_n]$$

with $q_n \geq 2$.

Proof: Let $\mu \in \{1, 2, \ldots, 2r\}$; then from (6.1)

$$\frac{p}{\mu} = \frac{[q_1, q_2, \ldots, q_n]}{[q_2, \ldots, q_n]} \quad \text{and} \quad p = [q_1, q_2, \ldots, q_n] \tag{7.1}$$

with $q_n \geq 2$; these continuant representations of p are distinct since otherwise, from (7.1), $p/\mu = p/\mu'$ for $\mu \neq \mu'$.

Let p have a representation $p = [q_1, q_2, \ldots, q_n]$ with $q_n \geq 2$. If $n = 1$ then $q_1 = q_n = p$ and we can take $\mu = 1$ in (7.1). If $n \geq 2$ then, since $q_n \geq 2$, it follows from properties 1 and 3 of Lemma 1 that $[q_2, \ldots, q_n] \leq (p-1)/2$ so that in (7.1) we have $[q_2, \ldots, q_n] \in \{2, \ldots, 2r\}$. ■

For $p = 13$ we obtain six continuant representations

$$13 = [13] = [6, 2] = [4, 3] = [3, 4] = [2, 6]$$

and

$$13 = [2, 1, 1, 2] = \begin{vmatrix} 2 & 1 & 0 & 0 \\ -1 & 1 & 1 & 0 \\ 0 & -1 & 1 & 1 \\ 0 & 0 & -1 & 2 \end{vmatrix}.$$

8. PROOF OF THEOREM 2. We begin with

Lemma 3. *Given any $n \in \mathbb{N}$ with $n \geq 2$ and any set of positive integers $\{q_1, q_2, \ldots, q_n\}$ define*

$$\begin{aligned} I_n(q_1, q_2, \ldots, q_n) := {} & [q_1, q_2, \ldots, q_n][q_2, q_3, \ldots, q_{n-1}] \\ & - [q_1, q_2, \ldots, q_{n-1}][q_2, \ldots, q_n]. \end{aligned} \tag{8.1}$$

Then

$$I_n(q_1, q_2, \ldots, q_n) = (-1)^n. \tag{8.2}$$

Proof: We have, from (4.2),

$$I_2(q_1, q_2) = [q_1, q_2] - [q_1][q_2] = 1. \tag{8.3}$$

For the general case we have, from property 6 of Lemma 1,

$$[q_1, \ldots, q_n] = [q_1, \ldots, q_{n-1}]q_n + [q_1, \ldots, q_{n-2}] \tag{8.4}$$

$$[q_2, \ldots, q_n] = [q_2, \ldots, q_{n-1}]q_n + [q_2, \ldots, q_{n-2}]. \tag{8.5}$$

Multiply (8.4) by $[q_2, \ldots, q_{n-1}]$ and (8.5) by $[q_1, \ldots, q_{n-1}]$ to give, using (8.1),

$$\begin{aligned} I_n(q_1, q_2, \ldots, q_n) &= [q_1, \ldots, q_{n-2}][q_2, \ldots, q_{n-1}] - [q_2, \ldots, q_{n-2}][q_1, \ldots, q_{n-1}] \\ &= -I_{n-1}(q_1, q_2, \ldots, q_{n-1}). \end{aligned}$$

Repeated application of this last result yields

$$I_n(q_1, q_2, \ldots, q_n) = (-1)^r I_{n-r}(q_1, \ldots, q_{n-r}) \quad (r \in \{1, 2, \ldots, n-2\});$$

taking $r = n - 2$ and using (8.3) gives $I_n(q_1, q_2, \ldots, q_n) = (-1)^{n-2} I_2 = (-1)^n$, so (8.2) follows, as required. ■

Proof of Theorem 2, Part 1. Let $p \in \mathbb{P}$ with $p \equiv 1 \pmod 4$. Then the proof of Theorem 1 ensures that p has at least one palindromic continuant representation

$$p = [q_1, \ldots, q_s, q_s, \ldots, q_1] \tag{8.6}$$

with $s \geq 1$ and $q_1 \geq 2$.

Define $x_0 \in \mathbb{N}$ by

$$x_0 := [q_2, \ldots, q_s, q_s, \ldots, q_1]; \tag{8.7}$$

it follows that, from $q_1 \geq 2$ and property 1 of Lemma 1,

$$1 < x_0 < (p-1)/2. \tag{8.8}$$

Now apply the result of Lemma 3 to the right-hand side of (8.6), with $n = 2s$, to obtain

$$[q_1, \ldots, q_s, q_s, \ldots, q_1][q_2, \ldots, q_s, q_s, \ldots, q_2]$$
$$-[q_1, \ldots, q_s, q_s, \ldots, q_2][q_2, \ldots, q_s, q_s, \ldots, q_1] = (-1)^{2s} = 1.$$

From (8.6), (8.7), and property 3 of Lemma 1 the preceding result reduces to

$$p[q_2, \ldots, q_s, q_s, \ldots, q_2] - x_0^2 = 1 \tag{8.9}$$

and so

$$x_0^2 \equiv -1 \pmod{p}. \tag{8.10}$$

This result, together with (8.8) completes the proof of Part 1. ■

Remarks

1. We note that (8.9) was known to Henry Smith and is stated at the end of [**20**].
2. We note also, from the proof of this part of Theorem 2, that the number

$$\lambda = x_0 \tag{8.11}$$

is a member of the set $\{2, 3, \ldots, r\}$ and, for this choice of λ, the quotient p/λ yields the palindromic continuant representation of p; see (6.5).

Examples

1. For the case when the prime $p = 13$ we have the following explicit results

$$p = [q_1, q_2, \ldots, q_s, q_s, \ldots, q_2, q_1] = [2, 1, 1, 2]$$
$$x_0 = [q_2, \ldots, q_s, q_s, \ldots, q_2, q_1] = [1, 1, 2] = 5$$
$$[q_2, \ldots, q_s, q_s, \ldots, q_2] = [1, 1] = 2.$$

The general result $1 < x_0 < (p-1)/2$ becomes $1 < 5 < 6$ in this case. We can now confirm the results (8.9) and (8.10):

$$p[q_2, \ldots, q_s, q_s, \ldots, q_2] - x_0^2 = 13 \cdot 2 - 5^2 = 1$$

and

$$x_0^2 = 25 = 26 - 1 \equiv -1 \pmod{13}.$$

2. Let $p = 1913$; then $p \equiv 1 \pmod{4}$ and, see (6.5) and (8.11), $\lambda = x_0 = 712$ since $712^2 = 506944 = 265 \times 1913 - 1 \equiv -1 \pmod{1913}$, with $1 < 712 < 956 = (1913 - 1)/2$. The Euclidean algorithm yields

$$\frac{p}{\lambda} = \frac{1913}{712} = \frac{[2, 1, 2, 5, 5, 2, 1, 2]}{[1, 2, 5, 5, 2, 1, 2]}$$

and since $[2, 1, 2, 5] = 43$, $[2, 1, 2] = 8$ we have $43^2 + 8^2 = 1849 + 64 = 1913$.

3. For $p = 969433$, with $p \equiv 1 \pmod{4}$ the appropriate palindromic continuant is $p = [2, 3, 4, 5, 6, 6, 5, 4, 3, 2]$ with $[2, 3, 4, 5, 6] = 972$ and $[2, 3, 4, 5] = 157$; thus $969433 = 972^2 + 157^2$.

4. For the example (1.1) produced by the Brillhart method, i.e. $p = 10^{50} + 577$, it may be shown that

$$\lambda = x_0 = 24574739597286316058804545812463447369459349571921$$

and the relevant palindromic continuant is

$$\begin{aligned} p = [&4, 14, 2, 4, 4, 1, 5, 1, 3, 4, 2, 17, 1, 1, 1, 3, 1, 2, 1, 7, 1, 4, 1, 1, 2, 7, 11, 1, \\ &1, 3, 3, 3, 1, 14, 1, 9, 1, 2, 1, 1, 1, 1, 22, 1, 1, 1, 1, 3, 21, 1, 1, 3, 3, 1, 5, 1, \\ &1, 5, 1, 3, 3, 1, 1, 21, 3, 1, 1, 1, 1, 22, 1, 1, 1, 1, 2, 1, 9, 1, 14, 1, 3, 3, 3, 1, \\ &1, 11, 7, 2, 1, 1, 4, 1, 7, 1, 2, 1, 3, 1, 1, 1, 17, 2, 4, 3, 1, 5, 1, 4, 4, 2, 14, 4], \end{aligned}$$

which has length 112. Thus we have, for the two squares representation,

$$\begin{aligned} u = [&4, 14, 2, 4, 4, 1, 5, 1, 3, 4, 2, 17, 1, 1, 1, 3, 1, 2, 1, 7, 1, 4, 1, 1, 2, 7, 11, 1, \\ &1, 3, 3, 3, 1, 14, 1, 9, 1, 2, 1, 1, 1, 1, 22, 1, 1, 1, 1, 3, 21, 1, 1, 3, 3, 1, 5, 1] \\ = &\ 761106534380835424545040 1 \end{aligned}$$

and

$$\begin{aligned} v = [&4, 14, 2, 4, 4, 1, 5, 1, 3, 4, 2, 17, 1, 1, 1, 3, 1, 2, 1, 7, 1, 4, 1, 1, 2, 7, 11, 1, \\ &1, 3, 3, 3, 1, 14, 1, 9, 1, 2, 1, 1, 1, 1, 22, 1, 1, 1, 1, 3, 21, 1, 1, 3, 3, 1, 5] \\ = &\ 648626890687392164224542 4 \end{aligned}$$

with $p = u^2 + v^2$, as in (1.1).

Proof of Theorem 2, Part 2. Suppose that r is another solution of the quadratic equation (2.1) with $r \neq x_0$ and $r \neq p - x_0$; without loss of generality we may suppose that r is a least, positive residue (mod p). Then $r^2 \equiv x_0^2 \equiv -1 \pmod{p}$ and hence p divides $r^2 - x_0^2 = (r - x_0)(r + x_0)$; since p is prime it divides $r - x_0$ or $r + x_0$. The former case implies $r \equiv x_0 \pmod{p}$, but since both r and x_0 are least, positive residues it follows that $r = x_0$. In the latter case $r \equiv -x_0 \equiv p - x_0 \pmod{p}$ and since r and $p - x_0$ are least, positive residues it follows that $r = p - x_0$. This contradiction completes the proof of Part 2. ■

9. THE "SMITH" PROOF OF THE GAUSS THEOREM. We are now in a position to give a proof, using the methods of Henry Smith, of the Gauss uniqueness result for the Fermat theorem, as presented in Theorem 1.

Proof: Let $p \in \mathbb{P}$ with $p \equiv 1 \pmod{4}$ and suppose that there are two, co-prime two squares representations: $p = u^2 + v^2$ and $p = s^2 + r^2$, with $u < v$, $s < r$.

Apply Algorithm 1 to the rational numbers v/u and r/s to obtain

$$1 < \frac{v}{u} = \frac{[q_1, q_2, \ldots, q_n]}{[q_2, \ldots, q_n]} \quad \text{and} \quad 1 < \frac{r}{s} = \frac{[t_1, t_2, \ldots, t_m]}{[t_2, \ldots, t_m]};$$

then

$$\begin{aligned} u &= [q_2, \ldots, q_n] \quad \text{and} \quad v = [q_1, q_2, \ldots, q_n] \\ s &= [t_2, \ldots, t_m] \quad \text{and} \quad r = [t_1, t_2, \ldots, t_m]. \end{aligned}$$

Hence, property 6 of Lemma 1 ensures that

$$\begin{aligned} p = u^2 + v^2 &= [q_2, \ldots, q_n]^2 + [q_1, q_2, \ldots, q_n]^2 \\ &= [q_n, \ldots, q_2, q_1, q_1, q_2, \ldots, q_n] \end{aligned} \tag{9.1}$$

and

$$p = s^2 + r^2 = [t_2, \ldots, t_m]^2 + [t_1, t_2, \ldots, t_m]^2$$
$$= [t_m, \ldots, t_2, t_1, t_1, t_2, \ldots, t_m]. \tag{9.2}$$

Part 1 of Theorem 2 guarantees that the continuants $[q_{n-1}, \ldots, q_1, q_1, \ldots, q_{n-1}, q_n]$ and $[t_{m-1}, \ldots, t_1, t_1, \ldots, t_{m-1}, t_m]$ are both solutions of the quadratic equation $x^2 \equiv -1 \pmod{p}$ and satisfy $1 < x < (p-1)/2$. From the uniqueness of this solution we have

$$[q_{n-1}, \ldots, q_1, q_1, \ldots, q_{n-1}, q_n] = [t_{m-1}, \ldots, t_1, t_1, \ldots, t_{m-1}, t_m] = \rho \text{ (say)}. \tag{9.3}$$

From (9.1), (9.2), and (9.3) it follows that

$$1 < \frac{p}{\rho} = \frac{[q_n, \ldots, q_2, q_1, q_1, q_2, \ldots, q_n]}{[q_{n-1}, \ldots, q_1, q_1, \ldots, q_{n-1}, q_n]} = \frac{[t_m, \ldots, t_2, t_1, t_1, t_2, \ldots, t_m]}{[t_{m-1}, \ldots, t_1, t_1, \ldots, t_{m-1}, t_m]}. \tag{9.4}$$

Applying Algorithm 1 to both the continuant terms in (9.4) shows that $m = n$ and $q_i = t_i$ $(i = 1, 2, \cdots, n)$; thus $u = s, v = t$ and the uniqueness result is established. ■

10. APPENDIX. In this appendix we reproduce the original 1855 paper [**20**] of Henry Smith.

DE COMPOSITIONE NUMERORUM PRIMORUM FORMAE $4\lambda + 1$ EX DUOBUS QUADRATIS

Sit

$$q_1 + \cfrac{1}{q_2 + \cfrac{1}{q_3 + \cfrac{}{\ddots \; + \cfrac{1}{q_n}}}}$$

fractio continua, cujus numerator, qui determinanti

$$\begin{vmatrix} q_1 & 1 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ -1 & q_2 & 1 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & -1 & q_3 & 1 & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & -1 & q_4 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 \\ 0 & 0 & 0 & 0 & \cdot & \cdot & -1 & q_n \end{vmatrix}$$

aequalis est, per hujusmodi formulam $[q_1\, q_2\, q_3 \ldots q_{n-1} q_n]$ exprimatur. Erit ergo

$$[q_1\, q_2 \ldots q_{i-1} q_i] = [q_i\, q_{i-1} \ldots q_2\, q_1]$$

et

$$[q_1 \ldots q_n] = [q_1\, q_2 \ldots q_i] \cdot [q_{i+1} \ldots q_n] + [q_1\, q_2 \ldots q_{i-1}] \cdot [q_{i+2} \ldots q_n];$$

quae aequationes pendent ab illa forma determinantali, ambae autem L. Eulero debentur.

Itaque, si quantitatum q *par* sumatur numerus, ipsaeque ita serie symmetrica disponantur, ut binae inter se aequales fiant, elucet, quantitatem $[q_1\, q_2 \ldots q_i\, q_i \ldots q_2\, q_1]$ summam fore duorum quadratorum inter se primorum; fit enim

$$[q_1\, q_2 \ldots q_i\, q_i \ldots q_2\, q_1] = [q_1\, q_2 \ldots q_i]^2 + [q_1\, q_2 \ldots q_{i-1}]^2.$$

Contra in numero quotientium *impari*, erit

$$[q_1 \ldots q_{i-1}\, q_i\, q_{i-1} \ldots q_2\, q_1] = [q_1 \ldots q_{i-1}] \cdot \{[q_1 \ldots q_i] + [q_1 \ldots q_{i-2}]\},$$

unde colligis, numerum $[q_1 \ldots q_i \ldots q_1]$ primum esse non posse, nec duplicem numeri primi; si quidem casus excipis, in quibus, aut i unitati aequatur, aut i binario, q_1 unitati.

Sit p numerus integer datus; $\mu_1, \mu_2, \ldots \mu_s$ series numerorum, qui ad p primi sunt, ipsiusque p dimidio minores.

Formentur fractiones continuae $\frac{p}{\mu_1}, \frac{p}{\mu_2}, \ldots, \frac{p}{\mu_s}$; quae omnes ita terminentur, ut is quotiens qui in extremo loco ponatur, unitatem superet. Hinc patet, quanta fuerit numerorum $\mu_1, \mu_2, \ldots \mu_s$ multitudo, tantum fore numerum determinantium $[q_1 \ldots q_n]$, qui dato numero p aequales erunt, neque praeter illos ullum dare ejusdem formae determinantem, cujus et primus et extremus quotiens unitate major sit, quique numero p aequalis esse possit.

Jam vero, quum duo determinantes $[q_1 \ldots q_n]$ et $[q_n \ldots q_1]$ aequales sint, quumque ipsum q_n unitate majus sit, apparet $[q_n \ldots q_1]$ ex una aliqua fractionum $\frac{p}{\mu}$ oriri. Unde sequitur, data quavis fractione $\frac{p}{\mu}$, inveniri posse aliam in eadem serie, quae quotientes eosdem, ordine inverso, repraesentet.

Sit p primus, formae $4\lambda + 1$; ut numerus determinantium ipsi p aequalium par existat. Quum ipse p unus e determinantium serie fiat, unus certo alius inveniri poterit in quo quotientium ordo invertendo non mutatur. Cum sit ergo

$$p = [q_1\, q_2 \ldots q_i\, q_i \ldots q_2\, q_1]$$

erit denique

$$p = [q_1\, q_2 \ldots q_i]^2 + [q_1\, q_2 \ldots q_{i-1}]^2.$$

Quam theorematis Fermatiani demonstrationem maxime elementarem esse patet, quum pendeat a conversione fractionum vulgarium in fractiones continuas.

Singulos autem formae $1 + x^2$ divisores ex duobus quadratis conflari, eodem modo demonstrare in promptu est. Sit enim

$$\mu\nu = 1 + x^2,$$

apparet fore

$$\mu = [q_1\, q_2 \ldots q_i\, q_i \ldots q_2\, q_1]$$

$$\nu = [q_2\, q_3 \ldots q_i\, q_i \ldots q_3\, q_2]$$

$$x = [q_1\, q_2 \ldots q_i\, q_i \ldots q_2].$$

Oxford, *Maio* 1854.

ACKNOWLEDGMENTS. Norrie Everitt thanks his three co-authors for their agreement to dedicate this paper to Paul Halmos who, from afar, has been his guide and mentor in mathematics. This paper should have been completed some years ago for a volume dedicated to Paul Halmos; apologies for the delay but I hope the paper is now the better for subsequent collaboration and extension.

All four authors thank Keith Hannabuss, Fellow and Tutor in Mathematics of Balliol College in Oxford, for his contribution to the Section on the life of Henry Smith; we have been guided by and quoted from his papers [**10**] and [**11**]; additionally we have had access to, and quoted from a yet unpublished account of the life of Henry Smith.

REFERENCES

1. C. W. Barnes, The representation of primes of the form $4n + 1$ as the sum of two squares, *Enseign. Math.* (2) **18** (1972) 289–299.
2. E. T. Bell, *Men of Mathematics*, Victor Gollancz Ltd., London, 1937.
3. J. Brillhart, Note on representing a prime as a sum of two squares, *Math. Comp.* **26** (1972), 1011–1013.
4. D. M. Burton, *Elementary Number Theory*, The McGraw- Hill Companies, Inc., New York, 1998.
5. G. E. Chrystal, *Algebra*: **I** and **II**, Adam and Charles Black, Edinburgh, 1889. The 6th. edition reprinted by Chelsea Publishing Co., New York, 1959.
6. H. A. Davenport, *The Higher Arithmetic*, Hutchinson House, London, 1952.
7. L. E. Dickson, *History of the Theory of Numbers* **11**, Chelsea Publishing Co., New York, 1966.
8. U. Dudley, *Elementary Number Theory*, 2nd. edition, W. H. Freeman and Company, New York, 1978.
9. J. A. Ewell, A simple proof of Fermat's two-square theorem, *Amer. Math. Monthly* **90** (1983) 635–637.
10. K. Hannabuss, The mathematician the world forgot, *New Scientist* **97** (1983) 901–903.
11. K. Hannabuss, Forgotten fractals, *Math. Intelligencer* **18** (1996) 28–31.
12. G. H. Hardy, *A Mathematician's Apology*, Cambridge University Press, 1969.
13. G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, Cambridge University Press, 1952.
14. G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, 5th edition, Oxford University Press, 1979.
15. T. Hawkins, *Lebesgue's Theory of Integration; Its Origins and Development*, Chelsea Publishing Co., New York, 1975.
16. D. R. Heath-Brown, Fermat's two-squares theorem, *Invariant.* (1984) 3–5.
17. C. Hermite, Note au sujet de l'article précedent, *J. Math. Pures Appl.* **13** (1848) 15.
18. T. Muir and W. H. Metzler, *A Treatise on the Theory of Determinants*, Dover Publications Inc., New York, 1960.
19. J.-A. Serret, Sur un théorème relatif aux nombres entiers, *J. Math. Pures Appl.* **13** (1848) 12–14.
20. H. J. S. Smith, De Compositione Numerorum Primorum $4\lambda + 1$ Ex Duobus Quadratis, *Crelle's Journal* **L** (1855) 91–92.
21. H. J. S. Smith, *The Collected Mathematical Papers of Henry John Stephen Smith*: **I** and **II**, (Edited by J.W.L. Glaisher), The Clarendon Press, Oxford, 1894. Reprinted by Chelsea Publishing Co., New York, 1965.
22. S. Wagon, The Euclidean algorithm strikes again, *Amer. Math. Monthly* **97** (1990) 125–129.
23. D. Zagier, A one-sentence proof that every prime $p \equiv 1 \pmod 4$ is a sum of two squares, *Amer. Math. Monthly* **197** (1990) 144.

FRANCIS CLARKE was educated at Birmingham University (B.Sc. 1968) and at the University of Warwick (M.Sc. 1969, Ph.D. 1971). Since 1971 he has taught in the Mathematics Department at the University of Wales Swansea. He has research interests both in algebraic topology and in number theory. The Bernoulli numbers are an enduring fascination that provide a link between the two fields. Other interests include playing the clarinet, hill walking, skiing, and sailing.
University of Wales Swansea, Singleton Park, Swansea SA2 8PP, Wales, UK.
f.clarke@swansea.ac.uk

W. NORRIE EVERITT was educated at the University of Birmingham, and then at Balliol College, Oxford in Great Britain. He obtained his D.Phil. from the University of Oxford in 1955 with thesis advisor Edward Charles Titchmarsh. He first heard of the Smith proof of the Fermat two squares theorem at a lecture given by the late Harold Davenport, in 1950, to the Invariant Society, University of Oxford. He is Professor Emeritus of the University of Birmingham.
School of Mathematics and Statistics, University of Birmingham, Edgbaston, Birmingham B15 2TT, England, UK
w.n.everitt@bham.ac.uk

LANCE L. LITTLEJOHN received his B.Sc. in 1975, his M.A. in 1976 (both in mathematics from the University of Western Ontario), and his Ph.D. in mathematics from Penn State University in 1981. His first academic position was at the University of Texas at San Antonio; since 1983, he has been at Utah State University. He enjoys all types of rigorous mathematics; in particular, his research interests in differential equations, operator theory, and special functions. Most of his non-mathematical time is spent with his wife, Wendy, and their two children, Alex and Mary (although he does manage to sneak some time to follow his beloved, but hapless, Detroit Tigers).
Department of Mathematics and Statistics, Utah State University, Logan, UT 84322-3900, USA
lance@math.usu.edu

ROELOF VORSTER received a B.Sc. from the University of Pretoria, South Africa in 1962 and started his career with a big bang as an explosives chemist at the world's largest dynamite factory. Realizing that mathematics is his first love, he joined the staff of the distance teaching University of South Africa (UNISA) where he obtained a Ph.D. in topology and category theory in 1972, and where he has been head of the mathematics department since 1990. It was during a visit to UNISA by Norrie Everitt and Lance Littlejohn in 1990 that the former delivered a fascinating talk that aroused great interest in the work of HJS Smith and the Fermat Two Squares Theorem.
Department of Mathematics, Applied Mathematics and Astronomy, University of South Africa, P.O. Box 392, 0001 Pretoria, Republic of South Africa
vorstsjr@alpha.unisa.ac.za

NOTES

Edited by **Jimmie D. Lawson and William Adkins**

Number Theory and Semigroups of Intermediate Growth

Melvyn B. Nathanson

The semigroup S is *finitely generated* if it contains a finite subset A such that every element of S can be written as a product of not necessarily distinct elements of A. The *length* of an element $s \in S$ with respect to A, denoted $l_A(s)$, is the smallest integer m such that s can be written as the product of m elements of A. The length of the identity element is 0. The *growth function* $\gamma_A(n)$ of S with respect to A counts the number of elements $s \in S$ of length at most n, that is, $\gamma_A(n) = \text{card}\{s \in S : l_A(s) \leq n\}$.

The semigroup S has *polynomial growth* of degree k with respect to A if $\gamma_A(n) \leq c_0 n^k$ for some positive constant c_0 and all sufficiently large n. For example, if S is the free abelian semigroup of rank k generated by a set A of cardinality k, then the number of elements of length exactly m is $\binom{m+k-1}{k-1}$, and

$$\gamma_A(n) = \sum_{m=0}^{n} \binom{m+k-1}{k-1} = \binom{n+k}{k} = \frac{n^k}{k!} + O(n^{k-1}),$$

so S has polynomial growth of degree k.

The semigroup S has *exponential growth* with respect to A if there exists a real number $\theta > 1$ such that $\gamma_A(n) \geq \theta^n$ for all sufficiently large n. For example, if S is the free semigroup of rank $k \geq 2$ generated by a set A of cardinality k, then the number of elements of length exactly m is k^m, and

$$\gamma_A(n) = \sum_{m=0}^{n} k^m = \frac{k^{n+1} - 1}{k - 1} > k^n,$$

so S has exponential growth.

The semigroup S has *intermediate growth* with respect to A if, for every positive integer k and for every real number $\theta > 1$, we have

$$n^k < \gamma_A(n) < \theta^n \tag{1}$$

for all sufficiently large n. The purpose of this note is to give a simple and self-contained proof of the existence of semigroups of intermediate growth.

The rate of growth of a semigroup S is an intrinsic property of the semigroup, that is, the growth is independent of the choice of generating set. This is easy to prove, since if A and B are two generating sets for S, then each element of B can be written as a finite product of elements of A. Thus, there exists a constant $c > 0$ such that $l_A(b) \leq c$ for all $b \in B$. If $s \in S$ is a product of m elements of B, then s can be written as a product of at most cm elements of A, and so $l_A(s) \leq c l_B(s)$ for

all $s \in S$. Therefore, if $l_B(s) \leq n$, then $l_A(s) \leq cn$, and $\gamma_B(n) \leq \gamma_A(cn)$. Similarly, there exists a constant $c' > 0$ such that $\gamma_A(n) \leq \gamma_B(c'n)$. It follows that the functions $\gamma_A(n)$ and $\gamma_B(n)$ have the same growth rates (polynomial, exponential, or intermediate).

It is not obvious that semigroups of intermediate growth exist. In the analogous case of a group G generated by a finite set A, the length of an element $g \in G$ is the smallest number m such that g can be represented in the form

$$g = a_1^{\pm 1} \cdots a_m^{\pm 1},$$

where $a_1, \ldots, a_m \in A$. In a famous problem in this MONTHLY in 1968, Milnor [**5**] asked if there exist groups of intermediate growth. The question for groups was answered affirmatively by Grigorchuk [**1**, **2**, **4**]. Semigroups of intermediate growth have also been constructed, but the proofs that their growth functions satisfy (1) are complicated. The semigroup constructed in Theorem 1 is known, but the proof by means of elementary number theory is new. We prove the intermediate growth property using only Chebyshev's Theorem [**6**, Theorem 6.3] that $\pi(x)$, the number of primes up to x, satisfies the inequalities

$$\frac{c_1 x}{\log x} \leq \pi(x) \leq \frac{c_2 x}{\log x} \tag{2}$$

for all $x \geq 2$, and the following simple upper bound for the Hardy–Ramanujan partition function $p(n)$.

Lemma 1. *For every integer* $n \geq 2$, $p(n) < 2n^{2\sqrt{n}}$.

Proof: The function $p(n)$ counts the number of partitions of a positive integer n. A partition of n is a sequence of positive integers $u_1, \ldots, u_s$ such that $n = u_1 + u_2 + \cdots + u_s$ and $u_1 \geq u_2 \geq \cdots \geq u_s$. Associated to this partition is an array of points, called the Ferrers graph, consisting of u_1 points on the first row, u_2 points on the second row, $\ldots$, u_s points on the s–th row. For example, corresponding to the partition $19 = 7 + 5 + 4 + 2 + 1$ is the graph in Figure 1.

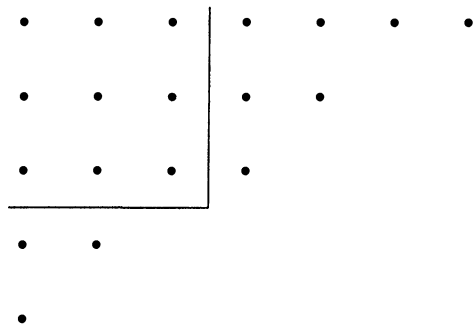


Figure 1

Consider the largest square array of dots that can be found in the upper left corner of the graph. In the example in Figure 1, this square consists of three lines, each with three points. The length of this square is some positive integer $r \leq \sqrt{n}$, and the square contains r^2 points. The remaining $n - r^2$ points in the graph must lie on the first r lines, to the right of the square, or on the first r columns, underneath the square. The number of ways to add these points to the graph is at most n^{2r}.

Therefore, the number of partitions of n satisfies

$$p(n) \leq \sum_{r=1}^{[\sqrt{n}]} n^{2r} < 2n^{2\sqrt{n}}.$$

Theorem 1. *Let S be the semigroup of 2×2 matrices generated by the set $A = \{a, b\}$, where*

$$a = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}.$$

Then

$$2^{2c_1\sqrt{n}/\log n} \leq \gamma_A(n) < n^2 p(n) < 2n^{2\sqrt{n}+2} \tag{3}$$

for some constant $c_1 > 0$ and all sufficiently large n. In particular, S has intermediate growth.

Proof: We observe that the matrices a and b satisfy the identities

$$b^2 = b,$$

$$a^k = \begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix},$$

and

$$ba^k b = \begin{pmatrix} k+1 & 0 \\ k+1 & 0 \end{pmatrix} = (k+1)b$$

for every positive integer k. Therefore, for any positive integers $k_1, \ldots, k_r$ we have

$$ba^{k_1}ba^{k_2}b \cdots ba^{k_r}b = \prod_{i=1}^{r} ba_i^{k_i}b = \left(\prod_{i=1}^{r}(k_i+1)\right)b. \tag{4}$$

Let $p_1, p_2, \ldots, p_r$ be distinct prime numbers not exceeding $\sqrt{n}$. Then

$$ba^{p_1-1}ba^{p_2-1}b \cdots ba^{p_r-1}b = \left(\prod_{i=1}^{r} p_i\right)b.$$

By Chebyshev's Theorem (2), the length of this element does not exceed

$$1 + \sum_{i=1}^{r} p_i \leq r\sqrt{n} + 1 \leq \sqrt{n}\,\pi(\sqrt{n}) + 1 \leq \frac{2c_2 n}{\log n} + 1 \leq n$$

for all sufficiently large n. Therefore, each of these elements is counted by the growth function $\gamma_A(n)$. These elements are distinct, since every positive integer is uniquely a product of primes, and so every subset of the primes up to $\sqrt{n}$ produces a different element of the semigroup S of length at most n. This gives the lower bound

$$\gamma_A(n) \geq 2^{\pi(\sqrt{n})} \geq 2^{2c_1\sqrt{n}/\log n}.$$

Next we compute an upper bound. Let $s \in S$ have length $l_A(s) \leq n$. There are three possibilities. First, there are $n+1$ elements of the form $s = a^u$ with $0 \leq u \leq n$. Second, there are $\binom{n+1}{2}$ elements of the form $s = a^u ba^v$ with $u, v \geq 0$ and $0 \leq u + v \leq n-1$. Third, we can have $s = a^u s' a^v$, where $s' = ba^{k_1}ba^{k_2}b \cdots ba^{k_r}b$, u, v are nonnegative integers with $u + v \leq n-3$, and $r, k_1, \ldots, k_r$ are positive integers such that $k_1 + \cdots + k_r + r + 1 \leq n$. Equation (4) implies that $s' = ba^{k_{\sigma(1)}}ba^{k_{\sigma(2)}}b \cdots ba^{k_{\sigma(r)}}b$ for every permutation σ of $1, \ldots, r$, so we can

assume that $k_1 \geq \cdots \geq k_r \geq 1$. Let $k_i' = k_i + 1$. We associate the following partition of n to the semigroup element s': $n = k_1' + \cdots + k_r' + 1 + \cdots + 1$, where $k_1' \geq \cdots \geq k_r' \geq 2$ and the number of 1's in the partition is

$$n - \sum_{i=1}^{r} k_i' \geq 1.$$

This is a one-to-one mapping of the elements s' to partitions of n. It follows that there are at most $p(n)$ semigroup elements s' of the form (4) with $k_1 + \cdots + k_r + r + 1 \leq n$. Since $u + v \leq n - 3$, there are $\binom{n-1}{2}$ choices of nonnegative integers u, v, and so there are at most $\binom{n-1}{2}p(n)$ semigroup elements of the third type. By Lemma 1, for $n \geq 2$ we have

$$\gamma_A(n) \leq n + 1 + \binom{n+1}{2} + \binom{n-1}{2}p(n) \leq n^2 p(n) < 2n^{2\sqrt{n}+2}.$$

This gives the upper bound. The semigroup S has intermediate growth since the left side of inequality (3) grows faster than any polynomial, while the right side grows slower than any exponential function. ■

Let A be a finite subset of a group, with $1 \in A$. Let S be the semigroup generated by A, and let G be the group generated by A. Denote by A^n the set of all products of n elements of A. Denote by $A^{\pm n}$ the set of all elements of the form $a_1^{\pm 1} \cdots a_n^{\pm 1}$. The growth function of the semigroup S with respect to A is $\gamma_{A,S}(n) = |A^n|$, and the growth function of the group G with respect to A is $\gamma_{A,G}(n) = |A^{\pm n}|$, Clearly, $\gamma_{A,S}(n) \leq \gamma_{A,G}(n)$ for all n. It is natural to ask if these two functions must have similar growth rates. Grigorchuk [3] proved that if $\gamma_{A,S}(n)$ has polynomial growth of degree k, then $\gamma_{A,G}(n)$ also has polynomial growth of degree k. It is not known if it is possible for the semigroup function $\gamma_{A,S}(n)$ to have intermediate growth while the group function $\gamma_{A,G}(n)$ has exponential growth.

REFERENCES

1. R. I. Grigorchuk, On Burnside's problem on periodic groups, *Funktsional. Anal. i Prilozhen.*, 14(1):53–54, 1980. English translation: *Functional Analysis Appl.* **14** (1980) 41–43.
2. R. I. Grigorchuk, On Milnor's problem of group growth, *Doklady Akad. Nauk SSSR*, 271(1):31–33, 1983. English translation: *Soviet Math. Dokl.* **28** (1983) 23–26.
3. R. I. Grigorchuk, Cancellation semigroups of polynomial growth, *Matem. Zamet.*, 43:305–319, 1988. English translation: Math. Notes (1988) 175–183.
4. R. I. Grigorchuk, On growth in group theory, in *Proceedings of the International Congress of Mathematicians, Kyoto, Japan*, 1990, Springer–Verlag, Tokyo, 1991, pp. 325–338.
5. J. Milnor, Problem 5603, *Amer. Math. Monthly* **75** (1968) 685–686.
6. M. B. Nathanson, *Additive Number Theory: The Classical Bases*, volume 164 of *Graduate Texts in Mathematics*, Springer-Verlag, New York, 1996.

Lehman College (CUNY), Bronx, NY 10468
nathansn@alpha.lehman.cuny.edu

The Gottschalk-Hedlund Theorem

Randall McCutcheon

In 1955 Gottschalk and Hedlund proved that if X is a compact metric space and $T: X \to X$ is a minimal homeomorphism (in other words, if the only closed sets $Y \subset X$ for which $TY \subset Y$ are $Y = X$ and $Y = \varnothing$), and if $f: X \to \mathbf{R}$ is continuous, then $f = g - Tg$ for a continuous function g (where $Tg(x) = g(Tx)$) if and only if there exists $K < \infty$ such that $|\sum_{n=0}^{N} T^n f(x)| < K$ for all $N \in \mathbf{N}$ and $x \in X$ [3].

Necessity is obviously violated if one allows X to be non-compact. However, the following theorem, which for Hausdorff spaces is due to Browder [1], is also true:

Theorem A. *Let (X, T) be minimal, where X is any topological space and $T: X \to X$ is continuous. Let $f: X \to \mathbf{R}$ be continuous and suppose that for some $K < \infty$ we have $|\sum_{n=0}^{N} T^n f(x)| < K$ for all $N \in \mathbf{N}$ and all $x \in X$. Then for some continuous g we have $f = g - Tg$.*

In 1993 the author, at the request of P. Schwartz, produced a proof of the Gottschalk-Hedlund theorem. It amounted to a minor alteration of a proof found by the author in 1989 of the fact (due to Bohr) that the integral of an almost periodic function, if it is bounded, is itself almost periodic; see [2, Theorem 5.2]. The proof, which sufficed for Theorem A as well, was so innocuous as to seem hardly interesting.

Schwartz, however, noticed something novel in the proof that "made the existence of a more general cocycle theorem seem likely" [6]. Indeed, he was able to adapt the proof to obtain a generalization of the Gottschalk-Hedlund theorem in a setting involving convolution operators. M. Lin and V. Bergelson then suggested that the proof would go in the context of Markov operators. Schwartz obtained something along these lines in [7].

Finally Lin and I. Kornfeld in [5] obtained a more general result of this type. Let X be compact space. A Markov operator on $C(X)$ is a positive contraction T with $T1 = 1$.

Theorem B. *Let X be a compact Hausdorff space, and let T be an irreducible Markov operator on $C(X)$ (see [5] for the definition of irreducible). If $g \in C(X)$ satisfies $\sup_N \|\sum_{j=0}^{N} T^j g\| < \infty$, then (and only then) there exists $f \in C(X)$ with $g = f - Tf$.*

The progressively more general results obtained in [6], [7], and [5] suggest that perhaps the proof we present here is somewhat interesting, after all (the central idea is mimicked in all three of these cases). Some form of the proof could, of course, be distilled from any of the aforementioned papers (and, in fact, it appears explicitly in [6]), but it wouldn't be completely clear how to do so most simply, as there are a few complications in the more general situations that require modification of the original argument.

Since no extra effort is involved, we actually prove a version of Theorem A that allows for continuous time, in which case the functional equation takes the form $-\frac{d}{dt} g(T_t x) = f(T_t x)$; see [4, Lemma 2.7]. Note that the proof remains valid for

maps and semiflows as well as homeomorphisms and flows. Finally, we derive as a corollary (Theorem D) the result that led to the proof's discovery: the integral of any almost periodic function over **R**, if bounded, is almost periodic.

Theorem C. *Let $(X,\{T_t\})$ be a minimal flow (with either discrete or continuous time). Let $f: X \to \mathbf{R}$ be continuous. If for all $x \in X$ there exists $K < \infty$ such that $|\int_0^L T_t f(x)\,dt| < K$ for all $L \geq 0$, then there exists a continuous function g such that $-\frac{d}{dt}g(T_t x) = f(T_t x)$ for all $x \in X$.*

Remark. In the case of discrete time, we take $\int_M^N$ to mean $\sum_M^{N-1}$.

Proof: For $x \in X$, put $g(x) = \sup_{N \geq 0} \int_0^N T_t f(x)\,dt$. We claim that for every $L \geq 0$, $g(x) = \sup_{N \geq L} \int_0^N T_t f(x)\,dt$.

Suppose the claim is false. Then there exists $x \in X$ and $L \geq 0$ such that $g(x) - \sup_{N \geq L} \int_0^N T_t f(x)\,dt = \epsilon > 0$. Pick $M < L$ such that $g(x) = \int_0^M T_t f(x)\,dt$. Then $\int_M^N T_t f(x)\,dt \leq -\epsilon$ for all $N \geq L$. Let $\delta = \inf_{S \geq L} \int_M^S T_t f(x)\,dt$ and fix $S \geq L$ with $\int_M^S T_t f(x)\,dt < \delta + \epsilon$.

By minimality of the flow, $\{T_t y : t \geq 0\}$ is dense in X for every $y \in X$ since its closure is a non-empty invariant set. In particular (taking $y = T_{L-M}x$), we may choose $r \geq L - M$ such that $T_r x$ lies in a neighborhood of x suitably chosen so as to ensure that

$$\int_{r+M}^{r+S} T_t f(x)\,dt = \int_M^S T_t f(T_r x)\,dt < \delta + \epsilon.$$

We have used continuity of the map $x \to \int_M^S T_t f(x)\,dt$; this is a fairly routine exercise. Hence

$$\int_M^{r+S} T_t f(x)\,dt = \int_M^{r+M} T_t f(x)\,dt + \int_{r+M}^{r+S} T_t f(x)\,dt < -\epsilon + \delta + \epsilon = \delta,$$

a contradiction that proves the claim.

We now have, for all $L \geq 0$,

$$g(T_L x) = \sup_{N \geq 0} \int_0^N T_t f(T_L x)\,dt = \sup_{N \geq 0} \int_L^{L+N} T_t f(x)\,dt = g(x) - \int_0^L T_t f(x)\,dt. \tag{1}$$

By the fundamental theorem of calculus, $-\frac{d}{dt}g(T_t x) = f(T_t x)$ (the discrete case follows directly from (1) by letting $L = 1$). All that remains, therefore, is to show that g is continuous.

By inspection g is lower semicontinuous. Let $h(x) = \inf_{N \geq 0} \int_0^N T_t f(x)\,dt$. Then

$$h(T_L x) = h(x) - \int_0^L T_t f(x)\,dt. \tag{2}$$

To see this, just replace f by $-f$ in the argument above. Clearly h is upper semi-continuous. Combining (1) and (2), we obtain $(g-h)(T_L x) = (g-h)(x)$ for all $L \geq 0$ and all $x \in X$. Let $x, y \in X$ and $\epsilon > 0$ be arbitrary. Notice that $(g-h)$ is lower semicontinuous. Utilizing the denseness of orbits, we obtain $L > 0$ such that $T_L x$ is "close enough" to y to ensure that $(g-h)(x) = (g-h)(T_L x) \geq (g-h)(y) - \epsilon$. However, since ϵ is arbitrary, $(g-h)(x) \geq (g-h)(y)$. Reversing the roles of x and y, we see that actually $(g-h)(x) = (g-h)(y)$, which implies

that $(g - h)$ is constant on X. Upper semicontinuity of g now follows from upper semicontinuity of h, since g differs from h by a constant. ■

Recall that a continuous function $f:[0,\infty) \to \mathbf{R}$ is *almost periodic* if for every $\epsilon > 0$ the set $\{n: \|f(x) - f(x+n)\|_u < \epsilon\}$ is *syndetic*, i.e. has bounded gaps. For $L \geq 0$, let $T_L f(x) = f(x+L)$. It may easily be shown that if f is almost periodic then $X = \overline{\{T_L f: L \geq 0\}}$ is compact in the uniform norm. Moreover $\{T_L\}_{L \geq 0}$ acts as a minimal isometric flow on X (with respect to the uniform norm) and $f(x)$ may be recovered by looking at the values at 0 of the members in the orbit of f: $f(x) = T_x f(0) = g(T_x f)$, where g is the (continuous) function that assigns to a member of X its value at 0.

More generally, if $(X, \{T_t\})$ is a minimal isometric flow on a compact space, $y \in X$ and $g: X \to R$ is continuous, the function $f(t) = g(T_t y)$ may be shown to be almost periodic. Hence the almost periodic functions are exactly those that arise (in this manner) from isometric flows on compact spaces. The following immediate corollary to Theorem C makes use of this fact.

Theorem D. *Let $F:[0,\infty) \to \mathbf{R}$ be almost periodic. If there exists $K < \infty$ such that $|\int_0^L F(t)\,dt| < K$ for all $L \geq 0$, then $H(s) = \int_0^s F(t)\,dt$ is almost periodic.*

Proof: We have an isometric flow on a compact space $(X, \{T_t\})$, a continuous function f on X, and a point $x \in X$ such that $F(t) = f(T_t x)$. The boundedness condition in the theorem is now exactly as in Theorem C. Hence there exists a continuous g on X such that $H(s) = \int_0^s T_t f(x)\,dt = g(x) - g(T_s x)$, which implies in particular that H is almost periodic. ■

REFERENCES

1. F. E. Browder, On the iterations of transformation in non-compact minimal dynamical systems, *Proc. Amer. Math. Soc.* **9** (1958) 773–780.
2. A. M. Fink, *Almost Periodic Differential Equations*, Lecture Notes in Mathematics, No. 377, Springer Verlag, New York, 1974.
3. W. H. Gottschalk and G. A. Hedlund, *Topological Dynamics*, AMS Colloquium Publications Vol. 36, American Mathematical Society, Providence, 1955.
4. R. Johnson, Minimal functions with unbounded integral, *Israel J. Math.* **31** (1978) 133–141.
5. I. Kornfeld and M. Lin, Coboundaries of irreducible Markov operators on $C(X)$, *Israel J. Math.* **97** (1997) 189–202.
6. P. Schwartz, A cocycle theorem with an application to Rosenthal sets, Ph.D. thesis, Ohio State University, 1994.
7. P. Schwartz, A cocycle theorem with an application to Rosenthal sets, *Proc. Amer. Math. Soc.* **124** (1996) 3689–3698.

University of Maryland, College Park MD 20742
randall@math.umd.edu

A Mean Value Theorem

Tadashi F. Tokieda

Several theorems go by this name. The present note adds to the assortment an unusual variant (Theorem 1), which involves the shape of the underlying region in an interesting way.

We work in Euclidean spaces, although Lemma 2 and the second inequality of Lemma 3 carry over to general Riemannian manifolds. ∇ and $|\ |$ denote gradient and norm with respect to the standard inner product $<\ ,\ >$, and ∂ stands for boundary. All our functions are real-valued. A *gradient curve* of a function f is an integral curve of ∇f.

Theorem 1. *Let f be a C^1-function on a closed ball B. Then there exists $b \in B$ at which $|\nabla f(b)| \cdot \operatorname{diam}(B) = \max f - \min f$.*

The proof is obtained via Lemmas 2 and 3.

Lemma 2. *Let f be a C^1-function without critical points on a compact region B. Then every gradient curve of f begins and ends on ∂B.*

Proof: Say a gradient curve $\gamma(s)$ is defined for s from s_- to s_+. We have

$$\lim_{s \to s_+} f(\gamma(s)) - \lim_{s \to s_-} f(\gamma(s)) = \int_\gamma \langle \nabla f,\ d\gamma \rangle$$

$$= \int_\gamma |\nabla f|\ |d\gamma| \quad \text{because } \gamma \text{ is tangent to } \nabla f \quad (*)$$

$$\geq \min|\nabla f| \cdot \operatorname{length}(\gamma).$$

On compact B, f is bounded, so if f has no critical points ($\min|\nabla f| > 0$), $(*)$ shows that $\operatorname{length}(\gamma)$ is finite and $\gamma(s_\pm)$ exist. Unless both $\gamma(s_-)$ and $\gamma(s_+)$ lie on ∂B, γ can be extended beyond s_- or s_+ by the existence theorem for solutions of differential equations, contradicting the choice of $s_\pm$. ∎

Remark. In Lemma 2, compactness is indispensable: think of the height function on an infinite vertical cylinder.

Lemma 3. *Let f be a C^1-function on a closed ball B. Then*

$$\min|\nabla f| \leq \frac{\max f - \min f}{\operatorname{diam}(B)} \leq \max|\nabla f|.$$

Proof: First inequality: If f has critical points on B, then $\min|\nabla f| = 0$. Otherwise consider the gradient curve γ *through the center of* B. γ reaches ∂B by Lemma 2, so that $\operatorname{length}(\gamma) \geq \operatorname{diam}(B)$; combine this with $(*)$ to get

$$\max f - \min f \geq \min|\nabla f| \cdot \operatorname{diam}(B).$$

Second inequality: Let l be a line segment that joins a minimum and a maximum of f. Since length$(l) \leq$ diam(B),

$$\max f - \min f = \int_l \langle \nabla f, dl \rangle \leq \int_l |\nabla f| \, |dl| \leq \max |\nabla f| \cdot \operatorname{diam}(B). \qquad \blacksquare$$

Remark. In Lemma 3, the first inequality is true only on a ball: $f(x, y) = x$ is a counterexample on $[0, 1] \times [0, 1]$. The second inequality holds on any convex region. Both become equalities for affine functions on balls.

Theorem 1 is now immediate:

Proof of Theorem 1 Apply Lemma 3 and the intermediate value theorem to $|\nabla f|$. $\blacksquare$

I do not know how close the mean value property of Theorem 1 comes to characterizing balls. However, Theorem 1 does admit a partial converse. To state it, we need a definition.

The *width* $w_B(\boldsymbol{e})$ of a compact region B in the direction of a unit vector $\boldsymbol{e}$ is defined as follows. 'Sandwich' B by a pair of parallel planes perpendicular to $\boldsymbol{e}$; $w_B(\boldsymbol{e})$ is the distance between these planes:

$$w_B(\boldsymbol{e}) = \max_{\boldsymbol{r} \in B} \langle \boldsymbol{e}, \boldsymbol{r} \rangle - \min_{\boldsymbol{r} \in B} \langle \boldsymbol{e}, \boldsymbol{r} \rangle.$$

B has *constant width* if $w_B(e)$ has the same value for all directions $\boldsymbol{e}$. A ball has constant width, but there are shapes of constant width that are not balls (e.g., Reuleaux's tetrahedron).

Aside. Why are lids on manholes round? Answer: because a lid whose rim is *not* a curve of constant width can fall into the hole if (un)suitably rotated. Of course, the lid and the hole need not be circular; any shape of constant width would be safe.

Return to the partial converse to Theorem 1.

Theorem 4. *Let B be a compact region such that for every linear function f on it, there exists $b \in B$ at which $|\nabla f(b)| \cdot \operatorname{diam}(B) = \max f - \min f$. Then B has constant width.*

Proof: Suppose B has maximal width in the direction of $\boldsymbol{e}_+$, minimal width in the direction of $\boldsymbol{e}_-$, and $w_B(\boldsymbol{e}_-) < w_B(\boldsymbol{e}_+)$. Then the linear function $f(\boldsymbol{r}) = \langle \boldsymbol{e}_-, \boldsymbol{r} \rangle$ violates the assumed property of f, as $|\nabla f| = 1$, $\operatorname{diam}(B) = w_B(\boldsymbol{e}_+)$, $\max f - \min f = w_B(\boldsymbol{e}_-)$. $\blacksquare$

University of Illinois, Urbana IL 61801
tokieda@math.uiuc.edu

On an Example of Jacobson

B. Sury

In Vol. III of Nathan Jacobson's celebrated book [**2**], there appears the following exercise on p. 49:

Let $\mathbb{F}_p$ be the field with p elements, and $P = \mathbb{F}_p(x, y)$ where x, y are indeterminates. Let E be the subfield $\mathbb{F}_p(x^p - x,\ y^p - x)$. Show that $[P : E] = p^2$, that P/E is not separable, and that P/E contains no purely inseparable element.

Now, it is seen immediately that Jacobson's example is really a nonexample. Surprisingly, none of the other standard graduate texts seem to give an example, although one can be found in [**1**, Ex. 17, Ch. V]. Here is another:

Example. Let $P = \mathbb{F}_p(x, y)$ and let E be the subfield $\mathbb{F}_p(x^p - x, y^p x)$. Then

(i) $[P : E] = p^2$,
(ii) P/E is not separable, and
(iii) P/E contains no purely inseparable element over E except those contained in E.

Recall that an element x in an algebraic closure $\overline{K}$ of a field K is *separable* if its minimal polynomial $f(T)$ in $K[T]$ has all roots (in $\overline{K}$) simple.

It is said to be *purely inseparable* over K if it is fixed by all K-automorphisms of $\overline{K}$. More generally, an algebraic extension L of K is said to be *purely inseparable* if the only elements of L that are separable over K are the elements of K itself. Any algebraic extension L of K is built in two stages: $K \subset L_{sep} \subset L$, where L_{sep} is separable over K, and L is purely inseparable over L_{sep}.

One has as a consequence of this definition:

Let $x \in \overline{K}$, and let $f(T)$ be its minimal polynomial over K. Then, the following statements are equivalent:

(*i*) *x is not separable over K*;
(*ii*) *The derivative $f'(T)$ is the zero polynomial*; *and*
(*iii*) *K is of characteristic $p > 0$, and $f(T) \in K[T^p]$.*

Under any of these equivalent hypotheses, if n is the smallest integer such that $x^{p^n} \in K$, then the minimal polynomial over K is $f(T) = T^{p^n} - x^{p^n}$.

We return to our example now.

$P = \mathbb{F}_p(x, y) \supset E = \mathbb{F}_p(a, b)$ where $a = x^p - x$, $b = y^p x$. Then, over E, y satisfies the polynomial $g(T) = T^{p^2} + rT^{p(p-1)} - s$, where $r = b/a$ *and* $s = b^p/a$. Also, $P = E(y)$.

We note:

(a) x is separable over E and y is inseparable over E.
The separability of x follows from the preceding remarks by looking at the polynomial $T^p - T - (x^p - x)$. This is a polynomial over E satisfied by x. In fact, the Artin-Schreier Theorem [**3**, Ch. 8] shows that this polynomial is irreducible and is the minimal polynomial of x over E. However, we do not need this fact for the proof.
The inseparability of y is a consequence of the observation that $T^p - y^p$ is the minimal polynomial of y over the field $E(x)$.

(b) y is not purely inseparable over E.
As $y^p \notin \in E$, one has also $y^{p^2} \notin \in E$; otherwise, from $g(y) = 0$ one concludes $y^{p(p-1)} \in E$, which would lead to the erroneous conclusion $y^p \in E$.

(c) x is not a p-th power in P.
This is easy to check by a simple comparison of like powers of x in view of the algebraic independence of x and y over $\mathbb{F}_p$.

Suppose $t \in P \setminus E$ is purely inseparable over E. Then $t^p \in E$ (because if $t^{p^n} \in E$ for some $n \geq 2$, then since the degree of $P = E(y)$ over E is at most p^2, $n \leq 2$. But, if $n \neq 1$, then P would be purely inseparable, a contradiction).

Look at $P \supset E(t) \supset E$. Now, the minimal polynomial of t over E is $T^p - t^p$, and $[E(t) : E] = p$. Note that $\alpha^p \in E$ for all $\alpha \in E(t)$.

Let $[P : E(t)] = l$, say. If the minimal polynomial of y over $E(t)$ is $f(T) = \sum_{i=0}^{l} a_i T^i$, then y satisfies the polynomial $f(T)^p = \sum_{i-0}^{l} a_i^p T^{ip}$. As this is the minimal polynomial of y over E, $f(T)^p$ divides $g(T)$. If $f(T)^p \sum u_i T^i = T^{p^2} + rT^{p(p-1)} - s$, one gets $u_i = 0$ if $i \not\equiv 0 \bmod p$. Renaming $u_{i/p}$ as v_i, the equation $\sum_{i=0}^{l} a_i^p T^{ip} \sum_{i=0}^{p-l} v_i T^{ip} = T^{p^2} + rT^{p(p-1)} - s$ gives inductively that $v_i = sb_i^p$ for some $b_i \in P$. Therefore, comparing the coefficients of $T^{p(p-1)}$ on both sides, we see $r = sv^p$ for some $v \in P$. This means that x is a p-th power in P, which is a contradiction.

Therefore, P has no purely inseparable elements outside of E.

Remarks. As a consequence of the proof, it is clear that $T^{p^2} + rT^{p(p-1)} - s$ is the minimal polynomial of y over E. The extension P of E is built up in two steps $P \supset E(x) \supset E$ with P purely inseparable of degree p over $E(x)$ and $E(x)$ separable of degree p over E.

REFERENCES

1. N. Bourbaki, *Algebre*, Actualites Scientifiques et Industrielles 1102, Hermann, Paris, 1950.
2. N. Jacobson, *Lectures in abstract algebra*, Vol. III, Van Nostrand, 1964.
3. S. Lang, *Algebra*, Addison-Wesley Publishing Company, Mass., 1965.

School of Mathematics, Tata Institute of Fundamental Research, Mumbai 400005, India.
sury@math.tifr.res.in

THE EVOLUTION OF...

Edited by **Abe Shenitzer**
Mathematics, York University, North York, Ontario M3J 1P3, Canada

Field Theory: From Equations to Axiomatization

Part I

Israel Kleiner

1. INTRODUCTION. The evolution of field theory spans a period of about 100 years, beginning in the early decades of the 19th century. This period also saw the development of the other major algebraic theories, namely group theory, ring theory, and linear algebra. The evolution of field theory was closely intertwined with that of the other three theories, as we shall see.

Abstract field theory emerged from three concrete theories—what came to be known as Galois theory, algebraic number theory, and algebraic geometry. These were founded, and began to flourish, in the 19th century. Of some influence in the rise of the abstract field concept were also the theory of congruences and (British) symbolical algebra. The 19th century's increased concern for rigor, generalization, and abstraction undoubtedly also had an impact on our story.

In this paper we discuss the sources of field theory as well as some of the main events in its evolution, culminating in Steinitz's abstract treatment of fields.

2. GALOIS THEORY. For three millennia (until the early 19th century) algebra meant solving polynomial equations, mainly of degrees up to 4. Field-theoretic ideas are implicit even here. For example, in solving the linear equation $ax + b = 0$, the four algebraic operations come into play and hence implicitly so does the notion of a field. In the case of the quadratic equation $ax^2 + bx + c = 0$, its solutions, $x = (-b \pm \sqrt{b^2 - 4ac})/2a$, require the adjunction of square roots to the field of coefficients of the equation. The concept of adjunction of an element to a field is fundamental in field theory.

Field-theoretic notions appear much more prominently, even if at first still implicitly, in the modern theory of solvability of polynomial equations. The groundwork was laid by Lagrange in 1770, but the field-theoretic elements of the subject were introduced by Abel and Galois in the early decades of the 19th century. Ruffini's 1799 proof of the insolvability of the quintic had a major gap because he lacked sufficient understanding of field-theoretic ideas [**16**].

Such ideas were starting points in Galois's 1831 "Mémoire sur les conditions de résolubilité des équations par radicaux" [**16**, p. 305]:

> One can agree to regard all rational functions of a certain number of determined quantities a priori. For example, one can choose a particular root

> of a whole number and regard as rational every rational function of this radical. When we agree to regard certain quantities as known in this manner, we shall say that we adjoin them to the equation to be resolved. We shall say that these quantities are adjoined to the equation. With these conventions, we shall call rational any quantity which can be expressed as a rational function of the coefficients of the equation and of a certain number of adjoined quantities arbitrarily agreed upon: . . . One can see, moreover, that the properties and the difficulties of an equation can be altogether different, depending on what quantities are adjoined to it.

It is clear that Galois has a good insight into the fields that we would denote today by $F(u_1, u_2, \ldots, u_n)$, obtained by adjoining the quantities $u_1, u_2, \ldots, u_n$ to the (field of) coefficients of an equation. In the specific example mentioned, he has in mind a quadratic field, $Q(\sqrt{d})$.

Galois was the first to use the term "adjoin" in a technical sense. The notion of adjoining the roots of an equation to the field of coefficients is central in his work [**9**], [**16**].

One of the fundamental theorems of the subject proved by Galois is the Primitive Element Theorem. This says (in our terminology) that if E is the splitting field of a polynomial $f(x)$ over a field F, then $E = F(V)$ for some rational function V of the roots of $f(x)$. Galois used this result to determine the Galois group of the equation $f(x) = 0$ [**1**], [**16**]. The Primitive Element Theorem was essential in all subsequent work in Galois theory until Artin bypassed it in the 1930s by reformulating Galois theory, for he felt that the theorem was not intrinsic to the subject [**9**].

3. ALGEBRAIC NUMBER THEORY. The central field-theoretic notion here, due independently to Dedekind and Kronecker, is that of an algebraic number field $Q(a)$, where a is an algebraic number. How did it arise? Mainly from three major number-theoretic problems: Fermat's Last Theorem (FLT), reciprocity laws, and representation of integers by binary quadratic forms. Although all three problems have to do with the domain of (ordinary) integers, in order to deal with them effectively it was found necessary to embed them in domains of what came to be known as algebraic integers. The following examples illustrate the ideas involved.

(a) To prove FLT for (say) $n = 3$, that is, to show that $x^3 + y^3 = z^3$ has no nonzero integer solutions, one factors the left side to obtain the equation $(x + y)(x + yw)(x + yw^2) = z^3$, where w is a primitive cube root of unity, $w = (-1 + \sqrt{3}\,i)/2$. This is now an equation in the domain $D = \{a + bw : a, b \in Z\}$ of algebraic integers. This approach to FLT (for $n = 3$) was essentially used by Euler and later by Lamé and others [**5**].

(b) Gauss's quadratic reciprocity law appeared in his *Disquisitiones Arithmeticae* of 1801. It says that $x^2 \equiv p \pmod{q}$ is solvable if and only if $x^2 \equiv q \pmod{p}$ is solvable, unless $p \equiv q \equiv 3 \pmod 4$, in which case $x^2 \equiv p \pmod{q}$ is solvable if and only if $x^2 \equiv q \pmod{p}$ is not. Here p and q are odd primes [**8**].

Gauss and others tried to extend this result to "higher" reciprocity laws. For example, for cubic reciprocity one asks about the relationship between the solvability of $x^3 \equiv p \pmod{q}$ and $x^3 \equiv q \pmod{p}$. These higher reciprocity-type problems are much more difficult to deal with than quadratic reciprocity. Gauss remarked

that [**8**, p. 108]:

> The previously accepted laws of arithmetic are not sufficient for the foundations of a general theory [of higher reciprocity]... Such a theory demands that the domain of arithmetic be endlessly enlarged.

His comments were no idle speculation. In fact, he himself began to implement the above "programme" by formulating and proving a law of *biquadratic reciprocity*. To do that he extended the domain of arithmetic by introducing what came to be known as the *gaussian integers* $G = \{a + bi : a, b \in Z\}$. He could not even formulate such a law without introducing G [**8**].

(c) The problem of representing integers by binary quadratic forms, namely determining when $n = ax^2 + bxy + cy^2$ $(a,'b,'c \in Z)$, goes back to Fermat. In particular, Fermat asked and answered the question: which integers n are sums of two squares, $n = x^2 + y^2$? In the *Disquisitiones* Gauss studied the *general* problem very thoroughly, developing a comprehensive and beautiful, but very difficult, theory. To gain a deeper understanding of Gauss's theory of binary quadratic forms, Dedekind found that he, too, needed to extend the domain Z of integers. For example, even in the simple case of representing integers as sums of two squares, it is the equation $(x + yi)(x - yi) = z^2$ rather than $x^2 + y^2 = z^2$ that yields conceptual insight [**1**], [**10**].

Dedekind's ideas. The fundamental question in extending the domain of ordinary arithmetic to "higher" domains is whether such domains behave like the integers, namely whether they are unique factorization domains (UFDs). It is this property that facilitates the solution of problems (a)–(c). While the domains D and G introduced above are UFDs, most domains that arise in connection with the three number-theoretic problems we have described are not. For example, when we factor the left side of $x^n + y^n = z^n$ for $n > 23$, the resulting domains are never UFDs. To rescue unique factorization in such domains Dedekind introduced (in Supplement X (1871) to Dirichlet's *Vorlesungen über Zahlentheorie*) ideals and prime ideals, and showed that every ideal in these domains is a unique product of prime ideals [**10**].

But what *are* the domains with restored unique factorization? To answer that—one of the fundamental questions of his theory—Dedekind needed to introduce fields, in particular *algebraic number fields* $Q(a)$, where a is a root of a polynomial with integer coefficients. These were the natural habitats of his domains, just as the rationals are the natural habitat of the integers. The domains in question were then defined as "the integers of $Q(a)$," namely those elements of $Q(a)$ that are roots of *monic* polynomials with integer coefficients. Dedekind showed that they form a commutative ring with identity and without zero divisors whose field of quotients is $Q(a)$ [**3**], [**10**], [**13**].

Given Dedekind's predisposition for abstraction—a rather rare phenomenon in the 1870s, he placed his theory in a broader context by giving axiomatic definitions of rings, fields, and ideals. Here is his definition of a field [**1**, p. 117]:

> By a field we will mean every infinite system of real or complex numbers so closed in itself and perfect that addition, subtraction, multiplication, and division of any two of these numbers again yields a number of the system.

To Dedekind, then, fields were subsets of the complex numbers, which is, of course, all he needed for his theory of algebraic numbers. Still, an axiomatic definition in number theory/algebra, even in this restricted sense, is remarkable for that time. Also remarkable are Dedekind's use of infinite sets ("systems"), which predates Cantor's, and his "descriptive" rather than "constructive" definition of a mathematical object as a set of all elements of a certain kind satisfying a number of properties.

The field concept was a unifying mathematical notion for Dedekind. Before his definition of a field he says [**4**, p. 131]:

> In the following paragraphs I have attempted to introduce the reader into a higher domain, in which algebra and number theory interconnect in the most intimate manner.... I became convinced that studying the algebraic relationship of numbers is most conveniently based on a concept that is directly connected with the simplest arithmetic properties. I had originally used the term "rational domain," which I later changed to "field."

Hilbert remarked that Gauss, Dirichlet, and Jacobi had also expressed their amazement at the close connection between number theory and algebra, on the grounds that these subjects have common roots in (as Dedekind would put it) the theory of fields [**4**].

Dedekind produced several editions of his groundbreaking theory of ideal decomposition in algebraic number fields. In his mature 1894 version (4th edition of Dirichlet's *Zahlentheorie*) he included important concepts and results on fields —nowadays standard—such as [**9**, pp. 130–132]:

(i) If S is any subset of the complex numbers containing the rationals, the intersection of all fields containing S is a field; it is called "rational with respect to S."

(ii) He defines field isomorphism, calling it "permutation of the field," as a mapping of a field E onto a field F that preserves all four operations of the field. He observes that if F is nonzero, the mapping is one-one. He also notes that the mapping is the identity on Q.

(iii) If E is a subfield of K, he defines the *degree* of K over E as the dimension of K considered as a vector space over E. He shows that if the degree is finite then every element of K is algebraic over E.

Kronecker's ideas. Kronecker's work was broader but much more difficult than Dedekind's. He developed his ideas over several decades, beginning in the 1850s, trying to frame a general theory that would subsume algebraic number theory and algebraic geometry as special cases. In his great 1882 work *Grundzüge einer arithmetischen Theorie der algebraischen Grössen* he developed algebraic number theory using an approach entirely different from Dedekind's. One of his central concepts was also that of a field—he called it "domain of rationality," defined as follows [**9**, p. 127]:

> The domain of rationality $(R', R'', R''', \ldots)$ contains every one of those quantities which are rational functions of the quantities $R', R'', R''', \ldots$ with integer coefficients.

Note how different Kronecker's "definition" of a field is from Dedekind's! It is a constructive description, rather than the kind of definition that would be accept-

able to us today. But it was dictated by Kronecker's views on the nature of mathematics.

Kronecker rejected irrational numbers as bona fide entities since they involve the mathematical infinite. For example, the algebraic number field $Q(\sqrt{2})$ was defined by Kronecker as the quotient field of the polynomial ring $Q[x]$ relative to the ideal generated by $x^2 - 2$, though he would have put it in terms of congruences rather than quotient rings. These ideas contain the germ of what came to be known as Kronecker's Theorem, namely that every polynomial over a field has a root in some extension field [**9**], [**13**].

It is interesting to compare this definition of $Q(\sqrt{2})$ with Cauchy's definition in the 1840s of the complex numbers as polynomials over the reals modulo $x^2 + 1$ (and compare the latter with Gauss's integers modulo p). Cauchy's rationale was to give an "algebraic" definition of complex numbers that would avoid the use of $\sqrt{-1}$.

Dedekind vs. Kronecker. Dedekind and Kronecker were great contemporary algebraists. Both published pathbreaking works on algebraic number theory. But their approaches to the subject were very different. Both were guided in their works by their "philosophies" of mathematics, and these too were very different [**13**]. Kronecker was perhaps the first preintuitionist, Dedekind likely the first preformalist (cf. Kronecker's "God made the [positive] integers, all the rest is the work of man" with Dedekind's "[The natural] numbers are a free creation of the human mind"). To Kronecker mathematics had to be constructive and finitary. Dedekind did not hesitate to use axiomatic notions and the infinite. While Kronecker made frequent pronouncements on these topics, Dedekind made few; his views became known mainly from his works—conceptual and abstract. Some examples:

(i) Since Kronecker's domains of rationality had to be generated by *finitely* many elements (the $R', R'', R''', \ldots$), his definition would not admit the totality of algebraic numbers as a field. Dedekind had no problem in considering the set of all complex numbers that are roots of polynomial equations with integer coefficients (viz. the set of all algebraic numbers) as a bona fide mathematical object.

(ii) On the other hand, Kronecker put no restriction on the nature of the entities $R', R'', R''', \ldots$—they could, for example, be indeterminates or roots of algebraic equations. So $Q(x)$ was a legitimate field to Kronecker. In fact, the adjunction of indeterminates to a field was a cornerstone of his approach to algebraic number theory. Dedekind, recall, defined his fields to be subsets of the complex numbers (but see Section 4).

(iii) Since Kronecker did not accept π (say) as a legitimate number, he identified $Q(\pi)$ with $Q(x)$ (x an indeterminate), thus claiming that transcendental numbers are indeterminate! To Dedekind $Q(\pi)$ was a perfectly legitimate entity not requiring any assistance from $Q(x)$.

4. ALGEBRAIC GEOMETRY. The examples of fields we have come across so far have been mainly fields of numbers. Here we encounter principally fields of functions, in particular, algebraic functions and rational functions. The ideas are due mainly to Kronecker and Dedekind-Weber.

Fields of algebraic functions. Algebraic geometry is the study of algebraic curves and their generalizations to higher dimensions, algebraic varieties. An *algebraic curve* is the set of roots of an algebraic function, that is, a function $y = f(x)$ defined implicitly by a polynomial equation $P(x, y) = 0$.

Several approaches were used in the study of algebraic curves, notably the analytic, the geometric-algebraic, and the algebraic-arithmetic. In the analytic approach, to which Riemann (in the 1850s) was the major contributor, the main objects of study were algebraic functions $f(w, z) = 0$ of a complex variable and their integrals, the so-called abelian integrals. It was in this connection that Riemann introduced the fundamental notion of a Riemann surface, on which algebraic functions become single-valued. Riemann's methods, however, were nonrigorous, relying heavily on the physically obvious but mathematically questionable Dirichlet Principle [**3**], [**11**].

Dedekind and Weber, in their important 1882 paper "Theorie der algebraischen Funktionen einer Veränderlichen," set for themselves the task of making Riemann's ideas rigorous, or, as they put it [**11**, p. 154]:

> The purpose of the[se] investigations ... is to justify the theory of algebraic functions of a single variable, which is one of the main achievements of Riemann's creative work, from a simple as well as rigorous and completely general viewpoint.

To accomplish this, they carried over to algebraic functions the ideas that Dedekind had earlier introduced for algebraic numbers. Specifically, just as an algebraic number field is a finite extension $Q(a)$ of the field Q of rational numbers, so Dedekind and Weber defined an *algebraic function field* as a finite extension $K = C(z)(w)$ of the field $C(z)$ of rational functions (in the indeterminate z). That is, w is a root of a polynomial $p(t) = a_0 + a_1 t + a_2 t^2 + \cdots + a_n t^n$, where $a_i \in C(z)$ (we can take $a_i \in C[z]$). Thus $w = f(z)$ is an algebraic function defined implicitly by the polynomial equation $P(z, w) = a_0 + a_1 w + a_2 w^2 + \cdots + a_n w^n = 0$. In fact, all the elements of $K = C(z)(w) = C(z, w)$ are algebraic functions.

Now let A be "the integers of K"; that is, A consists of the elements of $K = C(z)(w)$ that are roots of monic polynomials over $C[z]$ (cf. "the integers of $Q(a)$," Section 3). By analogy with the case of algebraic numbers, here too A is an integral domain and every nonzero ideal of A is a unique product of prime ideals [**1**], [**3**]. Incidentally, the meromorphic functions on a Riemann surface form a field of algebraic functions, with the entire functions as their "integers."

Dedekind and Weber were now ready to give a rigorous, algebraic definition of a Riemann surface S of the algebraic function field K: It is (in our terminology) the set of nontrivial discrete valuations on K. The finite points of S correspond to the ideals of A; to deal with points at infinity of S, they introduced the notions of "place" and "divisor" [**3**]. They developed many of Riemann's ideas on algebraic functions algebraically and rigorously. In particular, they gave a rigorous algebraic proof of the important Riemann-Roch Theorem [**1**], [**3**], [**11**].

Dedekind and Weber were at heart algebraists. They felt that algebraic function theory is intrinsically an algebraic subject, hence it ought to be developed algebraically. As they put it: "In this way, a well-delimited and relatively comprehensive part of the theory of algebraic functions is treated solely by means belonging to its own domain" [**11**, p. 156].

Beyond their technical achievements in putting major parts of Riemann's algebraic function theory on solid ground, the conceptual breakthrough by Dedekind and Weber lay in pointing to the strong analogy between algebraic number fields and algebraic function fields, hence between algebraic number theory and algebraic geometry. This analogy proved most fruitful for both theories.

Another noteworthy aspect of their work was its generality, in particular its applicability to arbitrary fields; see [**6**], [**15**].

Fields of rational functions. As noted earlier, algebraic geometry is the study of algebraic varieties. An algebraic variety is the set of points in R^n (or C^n) satisfying a system of polynomial equations $f_i(x_1, x_2, \ldots, x_n) = 0$, $i = 1, 2, \ldots, k$; the Hilbert basis theorem implies that finitely many equations will do. The ideal structure of the ring $R[x_1, \ldots, x_n]$ (or $C[x_1, \ldots, x_n]$) to which the polynomials $f_i(x_1, x_2, \ldots, x_n)$ belong is fundamental for the understanding of the algebraic variety, as is the "natural habitat" of that ring—its field of quotients $R(x_1, \ldots, x_n)$ (or $C(x_1, \ldots, x_n)$). These are the fields of (formally) *rational functions*. We have seen that such fields were also introduced by Kronecker in connection with his work in algebraic number theory [**6**], [**13**].

5. CONGRUENCES. Gauss introduced the congruence notation in the *Disquisitiones Arithmeticae* of 1801 and showed (among other things) that one can add, subtract, multiply, and divide congruences modulo a prime p, in effect that the integers modulo p form a field—a *finite* field of p elements. Inspired by Gauss's work on congruences, Galois introduced finite fields with p^n elements in an 1830 paper entitled "Sur la theorie des nombres."

Galois's aim was to study the congruence $F(x) \equiv 0 \pmod{p}$ as a generalization of Gauss's quadratic congruences (cf. Gauss's quadratic reciprocity law). Here $F(x)$ is a polynomial of degree n that is irreducible mod p, i.e., $F(x)$ is irreducible over the field Z_p. Galois showed that $F(x)$ has no integral roots [mod p]. His conclusion was that [**7**, pp. 277–278]:

> One should therefore regard the roots of this congruence as some kind of imaginary symbols..., symbols whose employment in calculation will often prove as useful as that of the imaginary $\sqrt{-1}$ in ordinary analysis.

He continues:

> Let i [an arbitrary symbol, *not* the complex number i] denote one of the roots of the congruence $F(x) \equiv 0$, which can be supposed to have degree n. Consider the general expression
>
> $$a + a_1 i + a_2 i^2 + \cdots + a_{n-1} i^{n-1}, \qquad (**)$$
>
> where $a, a_1, a_2, \ldots, a_{n-1}$ represent integers [mod p]. When these numbers are assigned all their possible values, expression (**) takes on p^n values, which possess, as I shall demonstrate, the same properties as the natural numbers in the *theory of residues of powers*.

Galois did, indeed, show that the expressions (**) form a field, now called a *Galois field*. He also showed that (in our terminology) the multiplicative group of that field is cyclic [**1**], [**7**], [**13**]. In an 1893 paper entitled "A doubly-infinite system of simple groups," E. H. Moore characterized the finite fields [**12**].

6. SYMBOLICAL ALGEBRA. In the third and fourth decades of the 19th century British mathematicians, notably Peacock, Gregory, and De Morgan, created what came to be known as symbolical algebra. Their aim was to set algebra—to them this meant the laws of operation with numbers, negative numbers especially—on

an equal footing with geometry by providing it with logical justification. They did this by distinguishing between *arithmetical algebra*—laws of operation with positive numbers, and *symbolical algebra*—a subject newly created by Peacock, which dealt with laws of operation with numbers in general.

Although the laws were carried over verbatim from those of arithmetical algebra, in accordance with the so-called Principle of Permanence of Equivalent Forms, the point of view was remarkably modern. Witness Peacock's definition of symbolical algebra, given in his *Treatise of Algebra* of 1830 [**14**, p. 35]:

> The science which treats of the combinations of arbitrary signs and symbols by means of defined though arbitrary laws.

Quite a statement for the early 19th century! Such sentiments were about a century ahead of their time. And of course one did have to wait about a century to have what Peacock had preached put fully into practice. Nevertheless, the creation of symbolical algebra was a significant development, even if not directly related to fields, signalling (according to some) the birth of abstract algebra [**2**].

REFERENCES

1. I. G. Bashmakova and E. I. Slavutin, Algebra and algebraic number theory, in *Mathematics of the 19th Century*, ed. by A. N. Kolmogorov and A. P. Yushkevich, Birkhäuser, 1992, pp. 35–135.
2. G. Birkhoff, Current trends in algebra, *Amer. Math. Monthly* **80** (1973) 760–782, and corrections in **81** (1974) 746.
3. N. Bourbaki, *Elements of the History of Mathematics*, Springer-Verlag, 1984.
4. L. Corry, *Modern Algebra and the Rise of Mathematical Structures*, Birkhäuser, 1996.
5. H. M. Edwards, *Fermat's Last Theorem: A Genetic Introduction to Algebraic Number Theory*, Springer-Verlag, 1977.
6. D. Eisenbud, *Commutative Algebra with a View Toward Algebraic Geometry*, Springer-Verlag, 1995.
7. E. Galois, Sur la theorie des nombres, English translation in S. Stahl, *Introductory Modern Algebra: A Historical Approach*, Wiley, 1997, pp. 277–284.
8. K. Ireland and M. Rosen, *A Classical Introduction to Modern Number Theory*, 2nd ed., Springer-Verlag, 1982.
9. B. M. Kiernan, The development of Galois theory from Lagrange to Artin, *Arch. Hist. Exact Sci.* **8** (1971/72) 40–54.
10. I. Kleiner, The roots of commutative algebra in algebraic number theory, *Math. Mag.* **68** (1995) 3–15.
11. D. Laugwitz, *Bernhard Riemann, 1826–1866*, Birkhäuser, 1999. Translated from the German by A. Shenitzer.
12. E. H. Moore, A doubly-infinite system of simple groups, *New York Math. Soc. Bull.* **3** (1893) 73–78.
13. W. Purkert, Zur Genesis des abstrakten Körperbegriffs I, II, *Naturwiss., Techn. u. Med.* **8** (1971) 23–37 and **10** (1973) 8–20. Unpublished English translation by A. Shenitzer.
14. H. M. Pycior, George Peacock and the British origins of symbolical algebra, *Historia Math.* **8** (1981) 23–45.
15. J. H. Silverman and J. Tate, *Rational Points on Elliptic Curves*, Springer-Verlag, 1992.
16. J.-P. Tignol, *Galois' Theory of Algebraic Equations*, Wiley, 1988.

PROBLEMS AND SOLUTIONS

Edited by **Gerald A. Edgar, Daniel H. Ullman,** and **Douglas B. West**

with the collaboration of Paul T. Bateman, Mario Benedicty, Paul Bracken, Duane M. Broline, Ezra A. Brown, Richard T. Bumby, Glenn G. Chappell, Randall Dougherty, Roger B. Eggleton, Ira M. Gessel, Bart Goddard, Jerrold R. Griggs, Douglas A. Hensley, John R. Isbell, Robert Israel, Kiran S. Kedlaya, Murray S. Klamkin, Fred Kochman, Frederick W. Luttmann, Vania Mascioni, Frank B. Miles, Richard Pfiefer, Cecil C. Rousseau, Leonard Smiley, John Henry Steelman, Kenneth Stolarsky, Richard Stong, Charles Vanden Eynden, and William E. Watkins.

Proposed problems and solutions should be sent in duplicate to the MONTHLY *problems address on the inside front cover. Submitted problems should include solutions and relevant references. Submitted solutions should arrive at that address before February 29, 2000; Additional information, such as generalizations and references, is welcome. The problem number and the solver's name and address should appear on each solution. An acknowledgement will be sent only if a mailing label is provided. An asterisk (*) after the number of a problem or a part of a problem indicates that no solution is currently available.*

Problem **10743** [1999; 586] in the June–July 1999 issue was misstated. Here is the corrected version.

10743. *Proposed by Călin Popescu, Université Catholique de Louvain, Louvain-La-Neuve, Belgium.* Let $R = \sum(-1)^i \binom{n}{i}$, where the sum is taken over all $i \in \{0, 1, \ldots, n-1\}$ such that $i+1$ is a quadratic residue modulo p, and let $N = \sum(-1)^j \binom{n}{j}$, where the sum is taken over all $j \in \{0, 1, \ldots, n-1\}$ such that $j+1$ is a quadratic nonresidue modulo p. Prove that exactly one of R and N is divisible by p.

PROBLEMS

10746. *Proposed by Stepan Tersian, University of Rousse, Rousse, Bulgaria.* Prove that

$$\int_0^\infty \left(e^{-y\sqrt{(s/x)^2+1}} - e^{-x\sqrt{(s/y)^2+1}}\right) \cos s \, ds = 0,$$

for all positive real numbers x and y.

10747. *Proposed by Athanasios Kalakos, Athens, Greece.* Find all differentiable functions $f : \mathbb{R} \to \mathbb{R}$ that are twice differentiable on an open interval containing 0, have exactly one real root, satisfy $f(1) = 1$, and satisfy $f'(f(t)) = 2f(t)$ for every $t \in \mathbb{R}$.

10748. *Proposed by Itshak Borosh, Douglas A. Hensley, and Joel Zinn, Texas A& M University, College Station, TX.* Let p and q be prime numbers, and let r be a positive integer such that $q \mid (p-1)$, $q \nmid r$, and $p > r^{q-1}$. Show that for any integers $a_1, a_2, \ldots, a_r$, if $\sum_{j=1}^r a_j^{(p-1)/q} \equiv 0 \bmod p$, then $\prod_{j=1}^r a_j \equiv 0 \bmod p$.

10749. *Proposed by Alain Grigis, Université Paris 13, Villetaneuse, France.* Let ABC be a triangle with a right angle at B and an angle of $\pi/6$ at A. Consider a billiard path in the triangle that begins at A, reflects successively off side BC at P, off side AC at Q, off side AB at R, off side AC at S, and then ends at B.

(**a**) Show that AP, QR, and SB are concurrent at a point X.

(**b**) Show that the angles formed at X measure $\pi/3$.

(**c**) Show that $AX = XP + PQ + QX = XR + RS + SX = 2XB$.

10750. *Proposed by Leonard Smiley, University of Alaska, Anchorage, AK.* For a positive integer m, express $\sum_{n=1}^{\infty}(n/\gcd(m,n))x^n$ as a rational function of x.

10751. *Proposed by Emeric Deutsch, Polytechnic University, Brooklyn, NY.* Let n be a positive integer, and let S_n be the set of all strings $a_1a_2\cdots a_n$ of positive integers satisfying $a_1 = 1$ and $a_{i+1} - a_i \in \{1, -1, -3, -5, \ldots\}$. For example, $S_5 = \{12345, 12343, 12341, 12323, 12321, 12123, 12121\}$. Find $|S_n|$.

10752. *Proposed by Gh. Costovici, Technical University "Gh. Asachi", Iasi, Romania.* For $n \in \mathbb{N}$, let a_n and b_n be complex numbers, with each $b_n \neq 0$. Let $s_n = a_1 + a_2 + \cdots + a_n$, and let $t_n = (1 - b_1/b_{n+1})\,a_1 + (1 - b_2/b_{n+1})\,a_2 + \cdots + (1 - b_n/b_{n+1})\,a_n$.
(**a**) Prove that if $\lim_{n\to\infty} b_{n+1}/b_n = 1$ and $\sum_{n=1}^{\infty}|s_n - t_n|^q$ converges for some $q \in (0, 1]$, then $\sum_{n=1}^{\infty} a_n$ converges.
(**b**) Prove that if $\sum_{n=1}^{\infty}|b_{n+1}/b_n - 1|^r$ and $\sum_{n=1}^{\infty}|s_n - t_n|^{r/(r-1)}$ converge for some $r \in (1, \infty)$, then $\sum_{n=1}^{\infty} a_n$ converges.

SOLUTIONS

A Zeta Function over a Recurrent Sequence

10486 [1995, 841]. *Proposed by Joseph H. Silverman, Brown University, Providence, RI.* Let $a, b > 0$ and $\alpha > 1$ be real numbers, and define $Z(s) = \sum_{n\in\mathbb{Z}}\left(a\alpha^n + b\alpha^{-n}\right)^{-s}$ for complex numbers s with positive real part.
(**a**) Prove that $Z(s)$ has a meromorphic continuation to all of $\mathbb{C}$.
(**b**) Find the poles of $Z(s)$.
(**c**) Find the residues of $Z(s)$ at its poles.

Solution I by David Bradley, University of Maine, Orono, ME. Let σ be the real part of s. Write

$$Z(s) = (a+b)^{-s} + \sum_{n=1}^{\infty}\left(a\alpha^n + b\alpha^{-n}\right)^{-s} + \sum_{n=1}^{\infty}\left(b\alpha^n + a\alpha^{-n}\right)^{-s}. \tag{1}$$

Without loss of generality, assume that $0 < a \le b$. We first consider the case $|\alpha| > \sqrt{b/a}$. We then have the two binomial expansions

$$\left(a\alpha^n + b\alpha^{-n}\right)^{-s} = \frac{a^{-s}\alpha^{-ns}}{\left(1 + ba^{-1}\alpha^{-2n}\right)^s} = a^{-s}\alpha^{-ns}\left(\sum_{k=0}^{m-1}\binom{-s}{k}\frac{b^k}{a^k}\alpha^{-2nk} + E_{m,n}(s)\right) \tag{2}$$

and

$$\left(b\alpha^n + a\alpha^{-n}\right)^{-s} = \frac{b^{-s}\alpha^{-ns}}{\left(1 + ab^{-1}\alpha^{-2n}\right)^s} = b^{-s}\alpha^{-ns}\left(\sum_{k=0}^{m-1}\binom{-s}{k}\frac{a^k}{b^k}\alpha^{-2nk} + F_{m,n}(s)\right), \tag{3}$$

where m is a fixed positive integer and $E_{m,n}(s) = O(\alpha^{-2mn})$ and $F_{m,n}(s) = O(\alpha^{-2mn})$. Since $|\alpha| > \sqrt{b/a}$, it follows from (1)–(3) that

$$\begin{aligned} Z(s) &= (a+b)^{-s} + \sum_{k=0}^{m-1}\binom{-s}{k}\left(\frac{b^k}{a^{s+k}} + \frac{a^k}{b^{s+k}}\right)\sum_{n=1}^{\infty}\alpha^{-n(s+2k)} + O\left(\sum_{n=1}^{\infty}\alpha^{-n(\sigma+2m)}\right) \\ &= (a+b)^{-s} + \sum_{k=0}^{m-1}\binom{-s}{k}\frac{a^{-s-k}b^k + b^{-s-k}a^k}{\alpha^{s+2k} - 1} + O\left(\sum_{n=1}^{\infty}\alpha^{-n(\sigma+2m)}\right). \end{aligned} \tag{4}$$

Since $E_{m,n}(s)$ and $F_{m,n}(s)$ are analytic for $\sigma > -2m$, it follows by analytic continuation that (4) is valid for $\sigma > -2m$. Since m is an arbitrary positive integer, we conclude that $Z(s)$ has a meromorphic continuation to the entire complex plane.

We now calculate the poles and residues from (4). For integer n, let $s_n = 2\pi i n/\log\alpha$. From (4), we see that all singularities of $Z(s)$ are simple poles, and these occur at points of the form $s_n - 2k$, where n is an integer and k is a nonnegative integer. The residue of $Z(s)$ at the pole $s_n - 2k$ is $\binom{2k-s_n}{k}a^k b^k(a^{-s_n} + b^{-s_n})/\log\alpha$.

If $(b/a)^{2n}$ is an odd power of α, then $s_n - 2k$ is an ordinary point, not a pole. Finally, although our derivation of (4) is valid only for complex numbers α with $|\alpha| > \sqrt{b/a}$, this restriction can be eased to $|\alpha| > 1$ by analytic continuation in α, provided that m remains fixed.

Solution II by Donald A. Darling, Newport Beach, CA. With $\beta = \log\alpha$, $\gamma = \frac{1}{2}\log(a/b)$, and $\delta = \sqrt{ab}$, the function $Z(s)$ takes the form

$$Z(s) = \frac{1}{(2\delta)^s}\sum_{n=-\infty}^{\infty}\frac{1}{\cosh^s(n\beta+\gamma)}.$$

Since the real part of s is positive, the function $\cosh^{-s}(n\beta+\gamma)$ satisfies the hypotheses of the Poisson summation formula $\sum_{n=-\infty}^{\infty} f(n) = \sum_{n=-\infty}^{\infty}\int_{-\infty}^{\infty} f(x)\cos(2\pi nx)dx$. Thus,

$$\begin{aligned}
&\sum_{n=-\infty}^{\infty}\frac{1}{\cosh^s(n\beta+\gamma)}\\
&= \sum_{n=-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\cos(2\pi nx)}{\cosh^s(\beta x+\gamma)}\,dx = \sum_{n=-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\cos(2\pi n(y-\gamma/\beta))}{\cosh^s(\beta y)}\,dy\\
&= \sum_{n=-\infty}^{\infty}\cos\left(\frac{2\pi n\gamma}{\beta}\right)\int_{-\infty}^{\infty}\frac{\cos(2\pi ny)}{\cosh^s(\beta y)}\,dy + \sum_{n=-\infty}^{\infty}\sin\left(\frac{2\pi n\gamma}{\beta}\right)\int_{-\infty}^{\infty}\frac{\sin(2\pi ny)}{\cosh^s(\beta y)}\,dy\\
&= \sum_{n=-\infty}^{\infty}\cos\left(\frac{2\pi n\gamma}{\beta}\right)\int_{-\infty}^{\infty}\frac{\cos(2\pi ny)}{\cosh^s(\beta y)}\,dy. \qquad (5)
\end{aligned}$$

Formula 3.985(1) (p. 540) of I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products*, Academic Press, 1980 states

$$\int_{-\infty}^{\infty}\frac{\cos(2\pi ny)}{\cosh^s(\beta y)}\,dy = \frac{2^{s-1}}{\beta\Gamma(s)}\,\Gamma\left(\frac{s}{2}+\frac{\pi in}{\beta}\right)\Gamma\left(\frac{s}{2}-\frac{\pi in}{\beta}\right).$$

With (5), this yields

$$\begin{aligned}
Z(s) &= \frac{1}{2\beta\Gamma(s)\delta^s}\sum_{n=-\infty}^{\infty}\cos\left(\frac{2\pi n\gamma}{\beta}\right)\Gamma\left(\frac{s}{2}+\frac{\pi in}{\beta}\right)\Gamma\left(\frac{s}{2}-\frac{\pi in}{\beta}\right)\\
&= \frac{\Gamma^2(s/2)}{2\beta\Gamma(s)\delta^s} + \frac{1}{\beta\Gamma(s)\delta^s}\sum_{n=1}^{\infty}\cos\left(\frac{2\pi n\gamma}{\beta}\right)\Gamma\left(\frac{s}{2}+\frac{\pi in}{\beta}\right)\Gamma\left(\frac{s}{2}-\frac{\pi in}{\beta}\right). \qquad (6)
\end{aligned}$$

By Stirling's formula for the gamma function in a vertical strip, the series in the last term of (6) converges uniformly in any vertical strip bounded away from the poles of the summands. Thus, $Z(s)$ has an analytic continuation to the entire complex s-plane, and the poles and residues may be calculated as in the first solution.

Editorial comment. When $|\alpha| > \sqrt{b/a}$, we can let $m \to \infty$ in (4) to obtain

$$Z(s) = (a+b)^{-s} + \sum_{k=0}^{\infty}\binom{-s}{k}\frac{a^{-s-k}b^k + b^{-s-k}a^k}{\alpha^{s+2k}-1},$$

which is valid for all complex s.

When $a = b = 1/2$ and $\alpha = e^{\pi z}$ with the real part of z positive, (6) yields $Z(s) = \sum_{n=-\infty}^{\infty}\cosh^{-s}(n\pi z)$. For positive integer s, such sums arise in the theory of elliptic

functions. For even s (C. B. Ling, On summation of series of hyperbolic functions, *SIAM J. Math. Anal.* 5 (1974) 551–561) and for all positive integral s (I. J. Zucker, The summation of series of hyperbolic functions, *SIAM J. Math. Anal.* 10 (1979) 192–206), $Z(s)$ can be evaluated in terms of elliptic functions. In particular, for $s = 1$, $0 < k < 1$, and $z = K(\sqrt{1-k^2})/K(k)$, we have $\sum_{n=-\infty}^{\infty} \cosh^{-1}(n\pi z) = (\sum_{n=-\infty}^{\infty} e^{-\pi n^2 z})^2 = (2/\pi)K(k)$, where $K(k)$ is the complete elliptic integral of the first kind (see B.C. Berndt, *Ramanujan's Notebooks, Part III*, Springer-Verlag, 1991, p. 102 and p. 138).

Solved also by D. Cantor, R. J. Chapman (U. K.), R. Holzsager, and the proposer.

A Matrix of Inequalities

10599 [1997, 566]. *Proposed by Fred Galvin, University of Kansas, Lawrence, KS.* Let $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$ be nonnegative numbers and let (a_{ij}) be an $m \times n$ matrix of nonnegative numbers with at least one nonzero entry in each row. Suppose that the inequality $\sum_{h=1}^{m} a_{hj}x_h \le \sum_{k=1}^{n} a_{ik}y_k$ holds whenever $a_{ij} > 0$. Show that $\sum_{i=1}^{m} x_i \le \sum_{j=1}^{n} y_j$.

Solution by Frank Jelen and Eberhard Triesch, Der Rheinisch-Westfälischen Technischen Hochschule, Aachen, Germany. Let A be the specified matrix, with columns $c_1, \ldots, c_n$. Let $x = (x_1, \ldots, x_m)^T$ and $y = (y_1, \ldots, y_n)^T$, and let $\mathbf{1}_k$ denote the column vector of length k with entries equal to 1.

Define $b = (b_1, \ldots, b_m)^T$ by $b_i = \max\{c_j^T x\colon a_{ij} > 0\}$; this is well-defined since each row contains a positive entry. Consider the linear programs

$$\text{minimize} \quad \mathbf{1}_n^T z \qquad \text{subject to} \quad Az \ge b \text{ and } z \ge 0 \tag{1}$$

and

$$\text{maximize} \quad b^T w \qquad \text{subject to} \quad A^T w \le \mathbf{1}_n \text{ and } w \ge 0. \tag{2}$$

These linear programs are duals of each other, and (1) has the feasible solution $z = y$. It thus suffices to show that there exists a feasible solution u of (2) with $b^T u \ge \mathbf{1}_m^T x$, since the Duality Theorem then yields $\mathbf{1}_n^T y \ge b^T u \ge \mathbf{1}_m^T x$.

Consider the nonnegative vector $u = (u_1, \ldots, u_m)^T$ defined by $u_i = x_i/b_i$ if $b_i > 0$ and $u_i = 0$ otherwise. Clearly $b^T u = \mathbf{1}_m^T x$.

For $1 \le j \le n$, define $I_j = \{i : a_{ij} > 0 \text{ and } x_i > 0\}$. For $i \in I_j$, we have $b_i \ge c_j^T x > 0$. Feasibility of u now follows from

$$c_j^T u = \sum_{i=1}^{m} a_{ij}u_i = \sum_{i \in I_j} a_{ij}\frac{x_i}{b_i} \le \frac{1}{c_j^T x}\sum_{i \in I_j} a_{ij}x_i = 1.$$

Solved also by the proposer.

A Complex Determinant

10601 [1997, 566]. *Proposed by Wen-Xiu Ma, Universität-GH Paderborn, Paderborn, Germany.* Let $n > 1$ be an integer and let $a_1, a_2, \ldots, a_n$ be complex numbers. Show that

$$\begin{vmatrix} 1 & a_1 & a_1^2 & \cdots & a_1^{2n-1} \\ 1 & a_2 & a_2^2 & \cdots & a_2^{2n-1} \\ \vdots & & & \ddots & \vdots \\ 1 & a_n & a_n^2 & \cdots & a_n^{2n-1} \\ 0 & 1 & 2a_1 & \cdots & (2n-1)a_1^{2n-2} \\ 0 & 1 & 2a_2 & \cdots & (2n-1)a_2^{2n-2} \\ \vdots & & & \ddots & \vdots \\ 0 & 1 & 2a_n & \cdots & (2n-1)a_n^{2n-2} \end{vmatrix} = (-1)^{n(n-1)/2} \prod_{1 \le i < j \le n} (a_i - a_j)^4.$$

Solution I by Robin J. Chapman, University of Exeter, Exeter, UK. Consider the Vandermonde matrix for $2n$ complex numbers $a_1, \ldots, a_{2n}$, in which the i, j-entry is a_i^{j-1}. The determinant is $\prod_{1 \le j < k \le 2n}(a_k - a_j)$. Subtracting row j from row $n+j$ turns row $n+j$ into

$$(a_{n+j} - a_j)\left[0, 1, a_{n+j} + a_j, \cdots, \sum_{r=0}^{2n-2} a_{n+j}^{2n-2-r} a_j^r\right].$$

These row operations do not change the determinant. When $a_{n+j} \neq a_j$ for each j, we may cancel $\prod_{j=1}^{n}(a_{n+j} - a_j)$ from the two expressions for the determinant to obtain

$$\begin{vmatrix} 1 & a_1 & a_1^2 & \cdots & a_1^{2n-1} \\ 1 & a_2 & a_2^2 & \cdots & a_2^{2n-1} \\ & \vdots & & \ddots & \vdots \\ 1 & a_n & a_n^2 & \cdots & a_n^{2n-1} \\ 0 & 1 & a_{n+1} + a_1 & \cdots & \sum_{r=0}^{2n-2} a_{n+1}^{2n-2-r} a_1^r \\ 0 & 1 & a_{n+2} + a_2 & \cdots & \sum_{r=0}^{2n-2} a_{n+2}^{2n-2-r} a_2^r \\ & \vdots & & \ddots & \vdots \\ 0 & 1 & a_{2n} + a_n & \cdots & \sum_{r=0}^{2n-2} a_{2n}^{2n-2-r} a_n^r \end{vmatrix} = \prod_{\substack{1 \le j < k \le 2n \\ k \neq n+j}} (a_k - a_j). \qquad (*)$$

By continuity, $(*)$ is also valid when $a_{n+j} = a_j$. Setting $a_j = a_{n+j}$ for each j in $(*)$ yields the desired result, since each difference $a_k - a_j$ for $j < k$ appears four times in the product on the right side of $(*)$, once in reverse order.

Solution II by Joseph J. Rushanan, The MITRE Corporation, Bedford, MA. We use the techniques from J. J. Rushanan, On the Vandermonde matrix, this MONTHLY **96** (1989) 921–924. Let A be the matrix whose determinant is given in the problem statement. Given a complex polynomial f defined by $f(z) = \sum_{i=0}^{2n-1} c_i z^i$, let $\mathbf{f} = [c_0, \ldots, c_{2n-1}]^T$. Then

$$A\mathbf{f} = \left[f(a_1), f(a_2), \ldots, f(a_n), f'(a_1), f'(a_2), \ldots, f'(a_n)\right]^T.$$

Let $f_k(z) = \prod_{i=1}^{k-1}(z - a_i)$ for $1 \le k \le n$, and let $f_k(z) = f_{k-n}(z)\prod_{i=1}^{n}(z - a_i)$ for $n+1 \le k \le 2n$. Since f_k is monic with degree $k-1$, the matrix $U = [\mathbf{f}_1 \cdots \mathbf{f}_{2n}]$ is upper-triangular with 1s on the diagonal. Furthermore, $L = AU$ is lower-triangular, since $f_k(a_j) = 0 = f'_{n+k}(a_j)$ if $1 \le j < k \le n$ and $f_{n+k}(a_j) = 0$ for all j.

Thus $\det A$ is the product of the diagonal terms of L, which are $f_k(a_k)$ and $f'_{n+k}(a_k)$ for $1 \le k \le n$. These terms consist only of factors of the form $(a_r - a_s)$ with $r \neq s$. A typical term $(a_s - a_r)$ with $r < s$ appears in $f_s(a_s)$ once, appears negated in $f'_{n+r}(a_r)$, and appears squared in $f'_{n+s}(a_s)$. This shows that A has the desired determinant.

The technique generalizes to higher derivatives.

Editorial comment. David Callan and Wai Wah Lau observed that generalizations involving higher derivatives have appeared in the literature, such as on page 400 of R. A. Horn and C. R. Johnson, *Topics in Matrix Algebra*, Cambridge Univ. Press, 1991. Several others noted that the formula holds for $a_1, \ldots, a_n$ in an arbitrary commutative ring. Indeed, every polynomial identity in $\mathbb{Z}[a_1, \ldots, a_n]$ holds over arbitrary commutative rings.

Solved also by M. Benedicty, J. C. Binz (Switzerland), G. L. Body (U. K.), D. Callan, L. L. Foster, J.-P. Grivaux (France), R. Holzsager, G. Keselman, N. Komanda, O. Kouba (Syria), W. W Lau, J. H. Lindsey II, G. R. Miller, M. McKee, J. H. Nieto (Venezuela), G. Peng, C. Popescu (Belgium), R. Richberg (Germany), J. H. Smith, P. Szeptycki, A. Tissier (France), J. Van hamme (Belgium), Wyoming Problems Circle, WMC Problems Group, and the proposer.

On a Convolution of Eulerian Numbers

10609 [1997, 664]. *Proposed by Donald E. Knuth, Stanford University, Stanford, CA.* Let $a(l,m,n) = \sum_{k=0}^{l}\binom{n}{k}(l+m-k)^{n-k}(k-l)^k$. Prove that $\sum_{l=1}^{n} a(l,m,n) = \big((m+n+1)/2\big)a(n,m,n) - \big((m+1)/2\big)m^n$.

Solution by David Callan, University of Wisconsin, Madison, WI. We compare coefficients of m^j to prove the desired identity. First we express $a(l,m,n)$ using Eulerian numbers. The classical Eulerian number $\left\langle {n \atop k} \right\rangle$ is the number of permutations of $[n] = \{1,\ldots,n\}$ consisting of k ascending runs (with descents at $k-1$ locations). These are counted by placing n numbered objects into k numbered boxes to avoid properties $P_1,\ldots,P_k$, where P_i is the property that box i is empty or has only objects greater than those in the preceding box. A careful application of the inclusion-exclusion principle yields the formula $\left\langle {n \atop k} \right\rangle = \sum_{j=0}^{k}(-1)^j\binom{n+1}{j}(k-j)^n$. Note that $\left\langle {0 \atop 0} \right\rangle = 1$.

The definition of $\left\langle {n \atop k} \right\rangle$ yields $\sum_{k=0}^{n}\left\langle {n \atop k} \right\rangle = n!$. We also need $\sum_{k=0}^{n} k\left\langle {n \atop k} \right\rangle = \frac{1}{2}(n+1)!$. To prove this combinatorially, we alter each permutation of $[n]$ with k runs by placing $n+1$ at the end of one run. This can be done in k ways and yields a permutation of $[n+1]$. A permutation of $[n+1]$ arises in this way if $n+1$ is at the end, not if $n+1$ is at the start, and otherwise if and only if the element preceding $n+1$ is greater than the element following it. Thus we obtain half the permutations of $[n+1]$.

We claim that

$$a(l,m,n) = \sum_{j=0}^{n}\sum_{i=0}^{l}\binom{n}{j}\left\langle {n-j \atop i} \right\rangle m^j, \qquad (*)$$

which we prove by comparing coefficients of m^j. By applying the binomial theorem to $(l-k+m)^{n-k}$, we extract the desired coefficient as $\sum_{k=0}^{l}(-1)^k\binom{n}{k}\binom{n-k}{j}(l-k)^{n-j}$. Rearranging binomial coefficients converts this to $\binom{n}{j}\sum_{k=0}^{l}(-1)^k\binom{n-j}{k}(l-k)^{n-j}$. After canceling the $\binom{n}{j}$, it remains only to rewrite the sum as $\sum_{i=0}^{l}\left\langle {n-j \atop i} \right\rangle$. We use our formula for Eulerian numbers, let $k = i-h$, apply the elementary identity $\sum_{h=0}^{r}(-1)^h\binom{n+1}{h} = (-1)^r\binom{n}{r}$, and finally interchange k with $l-k$ to obtain

$$\sum_{i=0}^{l}\left\langle {n-j \atop i} \right\rangle = \sum_{i=0}^{l}\sum_{h=0}^{i}(-1)^h\binom{n-j+1}{h}(i-h)^{n-j} = \sum_{k=0}^{l}k^{n-j}\sum_{h=0}^{l-k}(-1)^h\binom{n-j+1}{h}$$
$$= \sum_{k=0}^{l}(-1)^{l-k}\binom{n-j}{l-k}k^{n-j} = \sum_{k=0}^{l}(-1)^k\binom{n-j}{k}(l-k)^{n-j}.$$

Using $(*)$, we compute coefficients of m^j in the desired identity. For $j = n+1$, the contributions cancel. For $j = n$, the coefficient is n. For $j < n$, the coefficient of m^j in $\sum_{l=1}^{n} a(l,m,n)$ divided by $\binom{n}{j}$ is

$$\sum_{l=1}^{n}\sum_{i=1}^{l}\left\langle {n-j \atop i} \right\rangle = \sum_{i=1}^{l}(n+1-i)\sum_{i=1}^{n}\left\langle {n-j \atop i} \right\rangle = (n+1)\sum_{i=1}^{n}\left\langle {n-j \atop i} \right\rangle - \sum_{i=1}^{n} i\left\langle {n-j \atop i} \right\rangle$$
$$= (n+1)(n-j)! - \frac{1}{2}(n-j+1)! = (n-j)!\frac{n+j+1}{2}.$$

For $j < n$, the coefficient of m^j in $\big((m+n+1)/2\big)a(n,m,n) - \big((m+1)/2\big)m^n$ is

$$\frac{n+1}{2}\binom{n}{j}\sum_{i=1}^{n}\left\langle {n-j \atop i} \right\rangle + \frac{1}{2}\binom{n}{j-1}\sum_{i=1}^{n}\left\langle {n-j+1 \atop i} \right\rangle$$
$$= \frac{n+1}{2}\binom{n}{j}(n-j)! + \frac{1}{2}\binom{n}{j-1}(n-j+1)! = \binom{n}{j}(n-j)!\frac{n+j+1}{2}.$$

Editorial comment. From (*) we infer that $a(l, 1, n)$ is the total number of arrangements with at most l ascending runs that can be formed from subsets of $[n]$.

Solved also by R. J. Chapman (U. K.), Q. Darwish (Oman), H.-J. Seiffert (Germany), and the proposer.

A Variation on Additive Bases

10610 [1997, 664]. *Proposed by Richard Hall, University of Portsmouth, Portsmouth, England.* Given a positive integer m, let $C(m)$ be the greatest positive integer k such that, for some set S of m integers, every integer from 1 to k belongs to S or is a sum of two not necessarily distinct elements of S. For example, $C(3) = 8$ with $S = \{1, 3, 4\}$.

(a) Show that, for all $\epsilon > 0$, $1/4 < C(m)/m^2 < 1/2 + \epsilon$ for all sufficiently large m.

(b)* Improve the asymptotic bounds in part **(a)**.

Solution to **(a)** *by the National Security Agency Problems Group, Fort Meade, MD.* Let $[n]l$ denote the first n positive multiples of l. When m is even, with $m = 2t$, let $S = [t-1]1 \cup [t+1]t$. Since S has size m and represents all positive integers up to $(t+1)t + t$, we have $C(m) \geq t^2 + 2t$. Thus $C(m)/m^2 \geq (t^2 + 2t)/(2t)^2 > 1/4$.

When m is odd, with $m = 2t + 1$, let $S = [t]1 + [t+1](t+1)$. Since S has size m and represents all positive integers up to $(t+1)^2 + t + 1$, we have $C(m) \geq (t+1)(t+2)$. Thus $C(m)/m^2 \geq (t+1)(t+2)/(2t+1)^2 > 1/4$.

A set of size m represents at most $2m + \binom{m}{2}$ integers. Hence $C(m)/m^2 \leq 1/2 + 3/(2m) < 1/2 + \epsilon$ for $m > 3/(2\epsilon)$.

Solution to **(b)** *by the GCHQ Problems Group, Cheltenham, UK.* We show that $9/32 < C(m)/m^2 < 4/9 + \epsilon$ for all sufficiently large m.

For the lower bound, we construct a set that represents many integers by spreading the summands apart more quickly than in **(a)**. Write m as $16i + j$, where $-7 \leq j \leq 8$, and let $A = [1, 3i]$, $B = [2, 7i + j]3i$, $C = (7i + j)3i + [1, 3i](3i + 1)$, and $D = (7i + j)6i + 6i + [0, 3i]$, where $[x, y] = \{n \in \mathbb{Z}\colon x \leq n \leq y\}$. Let $S = A \cup B \cup C \cup D$.

From A and $A + A$ we get $[1, 6i]$, from $A + B$ we get $[6i + 1, (7i + j)3i + 3i]$, and from C and $A + C$ we get $[(7i + j)3i + 3i + 1, 30i^2 + 3i(j + 2)]$.

For $r \in [1, 4i + j]$ and $s \in [1, 3i]$, we have $(3i + r - s + 1)3i \in B$ and $(7i + j)3i + s(3i + 1) \in C$, and the sum of these integers is $(10i + r + j + 1)3i + s$. Thus $B + C$ contains $[(10i+2+j)3i+1, (14i+2j+1)3i+3i]$, which equals $[30i^2+3i(j+2)+1, (7i+j)6i+6i]$. Furthermore, $D \cup (A + D) \cup (B + D) = [(7i + j)6i + 6i, (7i + j)9i + 9i]$, and $C + D = [(7i+j)9i+9i+1, (7i+j)9i+3i(3i+1)+9i] = [(7i+j)9i+9i+1, 72i^2+i(12+9j)]$.

Since $|A \cup B \cup C \cup D| = m$, for large enough i we have

$$\frac{C(m)}{m^2} \geq \frac{72i^2 + i(12 + 9j)}{256i^2 + 32ij + j^2} = \frac{9}{32}\frac{8i^2 + i(j + 4/3)}{8i^2 + ij + j^2/32} > \frac{9}{32}.$$

To prove that $C(m)/m^2 < 4/9 + \epsilon$, we show that some of the $m + \binom{m}{2}$ pairs must be "wasted". This happens in two ways. First, the sum may be too big, as happens for any pair of numbers that both exceed $C(m)/2$. Second, note that $r - s = t - u$ if and only if $r + u = t + s$. Thus we obtain a wasted pair for each instance of identical differences.

Consider a set S that represents everything from 1 to μm^2, for some $\mu > 1/4$. We may assume that $S \subseteq [1, \mu m^2]$. Let $am = \left|S \cap [1, \mu m^2/2]\right|$. All pairs from the $(1 - a)m$ numbers above $\mu m^2/2$ are wasted. The smaller pairs have differences between 1 and $\mu m^2/2 - 1$, yielding wastage when $am + \binom{am}{2} > \mu m^2/2 - 1$.

Let $a \succeq b$ mean that $a > b - \epsilon$ for large enough m. Letting wm^2 be the number of wasted pairs, we have $w \succeq \max(0, (a^2 - \mu)/2) + (1 - a)^2/2$. Letting $f(a)$ denote this lower bound, we have $f'(a) = -(1 - a)$ for $a^2 < \mu$ and $f'(a) = 2a - 1$ for $a^2 > \mu$. The first quantity is negative and the second positive, since $\mu > 1/4$. Thus w is minimized at

$a^2 = \mu$, and hence $w \succeq (1 - \sqrt{\mu})^2/2$. Excluding the wasted sums yields $\mu \preceq 1/2 - w$, and so $2\mu \preceq 1 - (1 - \sqrt{\mu})^2 = 2\sqrt{\mu} - \mu$. Thus $\sqrt{\mu} \preceq 2/3$ and $\mu \preceq 4/9$.

Editorial comment. John Lindsey proved that $C(m)/m^2 < .499785077 + \epsilon$. Kevin Ford proved that $C(m)/m^2 < .48832 + \epsilon$ and conjectured that the number on the right can be replaced by $\pi/8 < .393$.

Solved also by R. J. Chapman, K. Ford, J. H. Lindsey II, and the proposer.

An Identity Involving Rooted Trees

10615 [1997, 767]. *Proposed by Joaquin Gómez Rey, Alcorcón, Madrid, Spain.* For n a positive integer, evaluate

$$\sum (k_1 + k_2 + \cdots + k_n)! \prod_{i=1}^{n} \frac{i^{(i-1)k_i}}{(k_i!)(i!)^{k_i}}$$

where the summation runs over all n-tuples $(k_1, k_2, \ldots, k_n)$ of nonnegative integers such that $k_1 + 2k_2 + \cdots + nk_n = n$.

Solution I by Anchorage Math Solutions Group, University of Alaska, Anchorage, AK. We show that the given expression a_n equals $n^n/n!$.

Let f, g, h be exponential generating functions with coefficients f_n, g_n, h_n, respectively. When $f(g(x)) = h(x)$, expanding the composition and collecting terms appropriately yields

$$h(x) = \sum_{n=0}^{\infty} n! \left(\sum_{l=1}^{n} f_l \cdot \sum \frac{\prod_{j=1}^{l} g_{\lambda_j}}{\prod_i (k_i!)(i!)^{k_i}} \right) \frac{x^n}{n!},$$

where the innermost sum is over partitions λ of n that have k_i parts of size i for each i and l parts all told. When $f_n = n!$ and $g_n = n^{n-1}$, this yields $h_n = n! a_n$.

To see that $h_n = n^n$, we use $g(x) = xe^{g(x)}$, a well-known consequence of the Lagrange Inversion Formula or of combinatorial manipulation of the exponential generating function for rooted trees. Since $f(x) = 1/(1 - x)$ as a formal series, $h(x) = 1/(1 - g(x))$. On the other hand, $\sum n^n x^n/n! = 1 + xg'(x)$. Thus it suffices to show that $(1 + xg'(x))(1 - g(x)) = 1$. This follows from computing $g'(x)$, which yields $xg'(x) = g(x) + xg'(x)g(x)$.

Solution II by GCHQ Problems Group, Cheltenham, U.K. We show that the sum equals $n^n/n!$ by counting a set of size n^n in another way, obtaining $n!$ times the desired sum. By Cayley's Formula, the set consisting of rooted trees on the vertex set $\{1, \ldots, n\}$ with one vertex marked has size n^n.

Alternatively, we view each such tree as a list of the subtrees that remain after deleting the edges on the path from the root to the marked vertex. First we partition the n points into subsets, using k_i subsets of size i for $1 \le i \le n$. For each n-tuple $(k_1, \ldots, k_n)$, there are $n!/\left(\prod_i i!^{k_i} k_i!\right)$ ways to do this is. We multiply this by $(\sum k_i)!$ and by $\prod_i (i^{i-1})^{k_i}$ to order the subsets and to place a rooted tree on the elements of each subset. When we assemble the trees, their roots in order form the path from the root to the marked vertex in the full tree. Since we obtain each rooted tree with marked vertex exactly once, we have proved that

$$\sum_{k_1, \ldots, k_n} \frac{n!}{\prod_i i!^{k_i} k_i!} \left(\sum_i k_i\right)! \prod_i \left(i^{i-1}\right)^{k_i} = n^n.$$

Editorial comment. The GCHQ Problems Group noted that omitting the factor $(\sum k_i)!$ yields unordered rooted forests on the n points, which correspond to rooted trees with $n + 1$

labeled points where point $n+1$ is constrained to be the root. Thus this approach also yields the identity

$$\sum_{k_1,\ldots,k_n} \prod_i \frac{(i^{i-1})^{k_i}}{i!^{k_i} k_i!} = \frac{(n+1)^{n-1}}{n!}.$$

Many solvers reduced the evaluation of the coefficient h_n in Solution 1 to the convolution $\sum_{k=0}^{n-1} \binom{n}{k} k^k (n-k)^{n-k-1}$ and obtained n^n for the sum by manipulating classical identites. The same formula is derived in F. Bergeron, G. Labelle, and P. Leroux, *Combinatorial Species and Tree-like Structures*, Cambridge Univ. Press, 1998, as an application of the pointing operation on the species of vertebrates, but it may be proved directly using the objects in Solution 2: Count the rooted trees with marked vertex having $n-k$ vertices in the marked subtree. When $k=0$, the root is the marked vertex, and n^{n-1} counts the rooted trees. When $k>0$, choose a set S of k vertices, a rooted tree with marked vertex on S, and a rooted tree on the remaining vertices. Let the marked vertex of the first tree be the parent of the root x in the second tree and view x as the marked vertex in the full tree.

Solved also by R. Bagby, N. Bansal (India), D. Beckwith, D. Callan, R. J. Chapman (U.K.), W. Chu (France), J. H. Lindsey II, H.-J. Seiffert (Germany), L. Takacs, D. Zeilberger, NSA Problems Group, and the proposer.

Divisors of Sums of Divisors

10617 [1997, 767]. *Proposed by James G. Merickel, Philadelphia, PA.* For a positive integer N, $\sigma(N)$ denotes the sum of the positive divisors of N. Given a positive integer n and a prime p, prove that there exist arbitrarily large sets S of multiples of n with the following property: For some positive integer m, the fraction $\sigma(N)/N$ reduces to a fraction whose denominator is p^m for every $N \in S$.

Solution by John P. Robertson, Berwyn, PA. Factor n as $p^s v$ with v relatively prime to p. Similarly factor $\sigma(v)$ as $p^t w$ with w relatively prime to p. We first consider the case when $p \neq 2$. By Dirichlet's Theorem, there are arbitrarily large sets of primes not congruent to -1 modulo p. Let Q be such a set not containing p or any primes that divide v. Let a be the product of the primes in Q.

Let $u = \phi(av(p-1))$, where ϕ denotes Euler's ϕ-function. Let $r = ku - 1$ with k large enough so that $r > \max\{s, t\}$, and let $m = r - t$. For each positive divisor b of a, the number $N = vp^r b$ is divisible by n because $r > s$. There are $2^{|Q|}$ such divisors b and hence there are $2^{|Q|}$ such N. Thus we need to show only that each fraction $\sigma(N)/N$, reduced, has denominator p^m.

Note that v, p^r, and b are pairwise relatively prime. Hence $\sigma(N)$ is $\sigma(v)\sigma(p^r)\sigma(b)$. Now $\sigma(p^r)$ is $(p^{r+1}-1)/(p-1)$. Because $r+1$ is a multiple of u and p is relatively prime to $av(p-1)$, we find that $(p^{r+1}-1)/(bv(p-1))$ is an integer that is not divisible by p. Also p does not divide the integer $w = \sigma(v)/p^t$, and p does not divide $\sigma(b)$, because no prime in Q is congruent to -1 modulo p. Thus

$$\frac{\sigma(N)}{N} = \frac{\left[\sigma(v)/p^t\right]\left[(p^{r+1}-1)/(bv(p-1))\right]\left[\sigma(b)\right]}{p^m},$$

where each item in brackets is an integer that is not divisible by p.

For the case $p = 2$, we repeat the argument with the following changes: Let Q be any set of odd primes that do not divide v. Use a^2 in place of a and b^2 in place of b, set $u = \phi(a^2 v(p-1))$, and set $N = vp^r b^2$. Since $\sigma(b^2)$ is odd, the argument proceeds in the same way.

Solved also by GCHQ Problems Group and the proposer.

REVIEWS

Edited by **Harold P. Boas**
Mathematics Department, Texas A & M University, College Station, TX 77843-3368

My Brain Is Open: The Mathematical Journeys of Paul Erdős. By Bruce Schechter. Simon& Schuster, 1998, 224 pp., $25.
The Man Who Loved Only Numbers: The Story of Paul Erdős and the Search for Mathematical Truth. By Paul Hoffman. Hyperion, 1998, 289 pp., $22.95.

Reviewed by **Albert A. Mullin**

Trying to understand the late, great Paul Erdős (1913–1996) is like trying to write a beautiful palindromic sonnet! Erdős was a genius wrapped in a mathematician inside an eccentric. A great problem poser and problem solver, he could see analogies between analogies. Personally independent in spite of his gregarious nature, he had an almost fanatical love of justice, partly due to some bad experiences in his own life. He always had a high regard for every form of intellectual aspiration and spiritual effort.

Like Mersenne, Erdős greatly advanced mathematics by extensively publicizing and propagating results and conjectures. Like Euler, Erdős published prodigiously and maintained his full intellectual faculties until the very day of his death. Since posthumous articles are still appearing, his complete works may reach a total of 1500 items. Like his fellow countryman John von Neumann, Erdős had a vivid interest in people. On the other hand, few mathematicians delighted in gossip like Johnny—certainly not Erdős. Like Gödel, Erdős had no formal students, although he wrote joint papers with nearly 500 different collaborators. Erdős had neither the "genius' ego" nor the "actor's ego," either of which can be disastrous for cooperation. Both Gödel and Erdős will likely have an endless number of students of their works.

Much has been made of Erdős' eccentricity; his peripatetic lifestyle; and his bantering use of terms such as "epsilon," "boss," and "slave" ("child," "wife," and "husband"), abbreviations such as P.G.O.M. (Poor Great Old Man) and S.F. (God, the Supreme Fascist), and metaphors such as THE BOOK (the S.F.'s transfinite volume containing the most elegant proofs of all theorems). Earlier this century, James Joyce made this whimsical, elliptical way of communicating intellectually respectable in his last two novels, *Ulysses* and *Finnegans Wake*. Erdős' eccentricity was often amusing and always harmless: sufficiently spectacular to provoke comment, but never serious enough to constitute disintegration. Edmund Landau summarized the situation well when he said, "Wir Mathematiker sind alle ein bißchen meschugge" ("We mathematicians are all a bit wacky").

There are numerous tales about Erdős; remarkably, many are true. Erdős himself would tell jokes in which he figured, but on close inspection, the jokes tell as much about others as they do about Erdős. For example, there is the neat story about Erdős traveling around the world by airplane, as he did so often. An epsilon is sitting between Erdős and the boy's mother. The boy is studying plane geometry, but he is puzzled by a certain result. Always looking for new talent, Erdős politely offers to help, but the boy refuses. The boy's mother scolds, "Listen to the man!" But the boy cries out to his mother, "What could *he* know about geometry?" Then there is the story about an old mathematician who declared that he had been

fortunate to have a wonderful life: he had proved elegant theorems and he had met many famous people. But he regretted never meeting Einstein. Erdős muttered something like, "Fear not, soon you will!" Erdős could joke positively about many situations that really were quite negative when viewed deeply enough. He enjoyed his job at Notre Dame: "Who could dislike a place with plus signs everywhere?"

In his later years, Erdős was constantly on the move. In Erdősian verse: "Another roof, another proof." His lifestyle after 1971 came straight from Keats: "What mad pursuit? What struggle to escape?" Erdős never married, perhaps because he preferred truth to beauty. He moved about freely and often, spreading the mathematical Word everywhere. He never sought "glory" because he realized it was a little like the elegant bed of Louis XIV: namely, magnificent, but full of bugs! Erdős lived slightly longer than Hilbert, but slightly shorter than Dedekind. His own wry comment on getting old was: "I am becoming stupider no more."

One must ask several questions about Erdős' super-productive life. How could he settle the Prime Number Theorem so easily without recourse to complex function theory? What prepared him for the necessary insights? Why didn't he settle the Riemann Hypothesis? Why didn't he settle Fermat's Last Theorem in an elementary way? I doubt that THE BOOK contains any 200-plus page proof of Fermat's Last Theorem, nor do I expect a short proof to be in the margin. Surely his deep conviction of the important role of discreteness guided Erdős away from various problems. His use of probabilistic methods in graph theory and Ramsey theory will long be remembered, but why didn't he find a *constructive* probabilistic method? Finally, one must be very puzzled that Erdős was rejected for work on the Manhattan Project at Los Alamos, while Klaus Fuchs was accepted. However, sometimes rejection is a great blessing in disguise.

Erdős' motto was: conjectures and proofs! His main interest was with the most primitive and basic objects of mathematics: the positive integers. Of course, he was interested in other areas of mathematics too: number theory; combinatorics (finite and transfinite); combinatorial geometry; set theory, but not logic; analysis; and probability theory, with special emphasis on probabilistic methods for the existence of formal objects. His approach to the investigation of this broad range of topics was uniform: unleash a barrage of cogent challenges. Then he would politely cajole people into working on them, almost always successfully. Most mathematicians are good at either conjecture or proof, but Erdős was great at both!

Here we have a pair of well written and sympathetic books (each authored by a non-mathematician) on Erdős' life and works. There is considerable overlap in topics between the books, and I could find no meaningful contradiction of views between them. Nevertheless, the two books do have different flavors and approaches. I found *My Brain Is Open* slightly more mathematical and more willing to go into details of technical results. On the other hand, I found *The Man Who Loved Only Numbers* more socially oriented; it even contains a detailed story about Erdős' intuition and Marilyn vos Savant's column on the Monty hall dilemma. It also has sub-biographies of several mathematicians who make more than cameo appearances. The different flavors of the two books are illustrated by the types of errors they contain. For instance, *The Man Who Loved Only Numbers* cites an erroneous value of π (pages 92 and 209), while *My Brain Is Open* confuses Ultra (a complex Allied operation involving intercepting and rendering intelligible enciphered enemy signals) with the codes used by Nazi Enigma machines. I suspect that life in Hungary between the World Wars was far worse than these books portray.

My Brain Is Open contains a broad spectrum of interesting topics: the many travels of the wandering mathematician; his early life and friendships in Hungary; mathematical proof (including THE BOOK where elegant proofs are found); squaring squares and Nazi Enigma machines; Ramsey theory; life at the Institute

for Advanced Study (including discussions with Einstein, Gödel, and von Neumann); number sieves; the great Prime Number Theorem controversy between Erdős and Selberg; and his many collaborations. *The Man Who Loved Only Numbers* also contains a rich vein of interesting subjects: straight from THE BOOK; problems with Sam and Joe; Einstein vs. Dostoevsky; marginal revenge; S.F. made the integers; and survivor's party. Each of the books has a splendid set of photographs of Erdős at various points in his life (only one photograph is common to the two books). Further, *The Man Who Loved Only Numbers* contains sample letters showing Erdős' noteworthy handwriting and terse style. Each of the bibliographies documents numerous technical results very well and points to useful sources for much additional information about Erdős.

In a world where the only modern mathematician that an average person can name is Theodore Kaczynski, it is wonderful to have these two accounts of the achievements and the humanity of the beloved Paul Erdős. I highly recommend *The Man Who Loved Only Numbers* to the intelligent general reader interested in the life and results of a mathematical genius, and I highly recommend *My Brain Is Open* to mathematicians, physicists, and computer scientists. For those saddened now that Erdős has, as he would have put it, "left," have heart, for now he knows whether the Riemann Hypothesis is true or not! And maybe the proof doesn't require any complex function theory.

506 Seaborn Drive NW, Huntsville, AL 35806 – 1831

TELEGRAPHIC REVIEWS

Edited by **Arnold Ostebee**
with the assistance of the Mathematics Departments of Carleton, Macalester, and St. Olaf Colleges

Telegraphic Reviews are designed to alert readers in a timely manner to new books appropriate to mathematics teaching and research. Special codes classify reviews by subject area and appropriate use:

T : Textbook	*P : Professional Reading*	*1–4: Semester*
C : Computer Software	*L : Undergraduate Library*	*** : Special Emphasis*
S : Supplementary Reading	*13: Grade Level*	*?? : Questionable*

Readers are advised that price information is subject to change. Selected books receive a second, more extensive review in the *Monthly*.

Books submitted for review should be sent to *Book Reviews Editor, American Mathematical Monthly, St. Olaf College, 1520 St. Olaf Avenue, Northfield, MN 55057-1098.*

Reference, P. *Handbook of Numerical Analysis, Volume VI.* Eds: P.G. Ciarlet, J.L. Lions. Elsevier Science, 1998, x + 689 pp, $164. [ISBN 0-444-82569-X] Three articles in two sections: "Numerical Methods for Solids (Part 3)" and "Numerical Methods for Fluids (Part 1)."

Mathematics Appreciation, S, L. *Strength in Numbers.* Sherman K. Stein. Wiley, 1996, xiii + 272 pp, $16.95 (P). [ISBN 0-471-32974-6] Subtitle (*Discovering the Joy and Power of Mathematics in Everyday Life*) is apt. Bits of real mathematics are described, explained, put in context, and admired. Requires only high school background; less in most parts. Clear, friendly, sometimes elegant; even the polemics are fun. PZ

Mathematics Appreciation, P, L. *Drawbridge Up: Mathematics—A Cultural Anathema.* Hans Magnus Enzensberger. Transl: Tom Artin. AK Peters, 1999, 48 pp, $5 (P). [ISBN 1-56881-099-7] A bilingual pamphlet (German original and English translation) containing the text of a talk given by poet and writer Enzensberger at the 1998 ICM meeting in Berlin. A lamentation over the "increasingly critical" paradox of the cultural isolation of mathematics at the peak of its "golden age" marked by "spectacular" achievements and applications. LAS

Recreational Mathematics, S, L*. *The Mathemagician and Pied Puzzler: A Collection in Tribute to Martin Gardner.* Eds: Elwyn Berlekamp, Tom Rodgers. AK Peters, 1999, x + 266 pp, $34 (P). [ISBN 1-56881-075-X] Proceedings of the first "gathering for Gardner" held in 1993. Games, puzzles, and recreational mathematics—all accessible to general readers—contributed by the world's foremost magicians, puzzlists, and mathematicians. LCL

History, S(14–17), P, L.** *Euler: The Master of Us All.* William Dunham. Dolciana Math. Expos., No. 22. MAA, 1999, xxviii + 185 pp, $29.95 (P). [ISBN 0-88385-328-0] A necessarily sparse sample of Euler's major contributions to diverse areas of mathematics (number theory, infinite series, complex variables, algebra, geometry, and combinatorics) conveyed in contemporary notation yet faithful to Euler's approach—even his occasional "mathematical madness." Each chapter sets the stage by describing clearly the efforts that preceded Euler, the magnitude of the challenge he faced, and the impact of the contribution he made. LAS

History, P. *Selected Publications of Eugene L. Lawler.* Eds: K. Aardal, *et al.* CWI Tract, V. 126. Centrum voor Wiskunde en Informatica, 1999, x + 318 pp, Dfl. 60 (P). [ISBN 90-6196-484-9] 26 of Lawler's technical and expository papers as well as a complete list of his publications.

Foundations, T(14), L. *An Introduction to Abstract Mathematics.* Robert J. Bond, William J. Keane. Brooks/Cole, 1999, xix + 323 pp. [ISBN 0-534-95950-7] Based on a "transitions" course at Boston College. Chapters 1–5 introduce logic, sets, functions, relations, and the integers. Chapters 6–8 give applications to infinite sets, real and complex numbers, polynomials. Year of calculus assumed. Many examples and exercises. JD

Foundations, S(13–18), P, L*. *A Mathemat-*

ical Mystery Tour: Discovering the Truth and Beauty of the Cosmos. A.K. Dewdney. Wiley, 1999, vi + 218 pp, $22.95. [ISBN 0-471-23847-3]. An imaginative, meandering tour of mathematics in the service of resolving two abiding mysteries: mathematics' "unreasonable" utility, and whether it is discovered or created. Dewdney suggests that perhaps the Pythagoreans were right after all—that mathematics exists, awaiting discovery, in an other-worldly "holos" out of which the cosmos comes into being. LAS

Discrete Mathematics, T(13–15: 1, 2). *Discrete Mathematics and Its Applications, Fourth Edition.* Kenneth H. Rosen. McGraw-Hill, 1999, xxii + 804 pp. [ISBN 0-07-289905-0] New section on generating functions; more on rules of inference; ties to new web site. (*Second Edition*, TR, November 1991.) DB

Number Theory, T(14–16: 1, 2). *Elements of the Theory of Numbers.* Joseph B. Dence, Thomas P. Dence. Academic Pr, 1999, xvii + 517 pp. [ISBN 0-12-209130-2] Well-done but fairly standard introduction to number theory. Includes introduction to number fields and partition theory. DB

Number Theory, T(13–15: 1), L. *The Mathematics of Ciphers: Number Theory and RSA Cryptography.* S.C. Coutinho. AK Peters, 1999, xv + 196 pp, $30. [ISBN 1-56881-082-2] A gentle amd very readable introduction to number theory that culminates in the RSA public-key cryptosystem. DB

Linear Algebra, T(14). *Introduction to Linear Algebra.* Donald J. Wright. McGraw-Hill, 1999, ix + 392 pp. [ISBN 0-07-072098-3] From the Preface: "The goal is to learn to think in terms of linear algebra notions ... and that sort of familiarity comes from using the ideas in a substantive way." Contains a variety of useful examples. PF

Ring Theory, P, L. *Rings and Things and a Fine Array of Twentieth Century Associative Algebra.* Carl Faith. Math. Surv. & Mono., V. 65. AMS, 1999, xxxii + 422 pp, $99. [ISBN 0-8218-0993-8] Exhaustive survey of 125 years of associative algebras, ring and module theory. The author's *Algebra I* and *II* serve as the foundation for this survey. Also includes the author's thoughts on mathematics and mathematicians of the last 50 years. Bibliography has over 1600 references. JD

Algebra, T(17: 2), P, L. *Post-Modern Algebra.* Jonathan D.H. Smith, Anna B. Romanowska. Pure & Appl. Math. Wiley, 1999, xi + 370 pp, $69.95. [ISBN 0-471-12738-8] Introduction to algebra from an applications-based perspective. Traditional topics of groups, rings, fields, modules are accompanied by monoids, quasigroups, lattices, Boolean algebras, and more. Structures unified by techniques of universal algebra and category theory. JD

Algebra, P. *Differential and Difference Dimension Polynomials.* M.V. Kondratieva, *et al.* Math. & Its Applic., V. 461. Kluwer Academic, 1999, xiii + 422 pp. [ISBN 0-7923-5484-2]

Real Analysis, T*(16–17: 3), L. *A Course in Real Analysis.* John N. McDonald, Neil A. Weiss. Academic Pr, 1999, xvii + 745 pp. [ISBN 0-12-742830-5] Each chapter begins with a biography of a key contributor (from Cantor to Daubechies). Discusses the usual topics: set theory, real number system, Lebesgue theory, metric spaces. Further interest is piqued by chapters on probability, harmonic analysis, dynamical systems. Lots of exercises. Teachers of analysis—have a look! KS

Real Analysis, T(17), P. *Real Analysis—With an Introduction to Wavelet Theory.* Satoru Igari. Transl: Satoru Igari. Transl. of Math. Mono., V. 177. AMS, 1998, xiii + 256 pp, $89. [ISBN 0-8218-0864-8] Covers Lebesgue measure and integration, differentiation, abstract measures, L^p-spaces, distribution theory, Fourier analysis. Rather dry "definition-theorem-proof" exposition. Introduction to wavelets is very brief. BH

Partial Differential Equations, T(16: 1), C, L. *Partial Differential Equations and Boundary Value Problems with Maple V.* George A. Articolo. Academic Pr, 1998, xii + 628 pp, (P), with CD-ROM. [ISBN 0-12-064475-4] Traditional approach to PDEs incorporating a brief discussion of useful Maple commands and many problems solved using Maple. Begins with a review of ODEs and a chapter on Sturm–Liouville eigenvalue problems. Includes discussion of heat, wave, and Laplace equations. PG

Partial Differential Equations, P. *Differential Operators and Spectral Theory: M. Sh. Birman's 70th Anniversary Collection.* Eds: V. Buslaev, M. Solomyak, D. Yafaev. AMS Transl., Ser. 2, V. 189. AMS, 1999, viii + 285 pp, $99. [ISBN 0-8218-1387-0]

Functional Analysis, P. *Introduction to the Theory and Applications of Functional Differential Equations.* V. Kolmanovskii, A. Myshkis. Math. & Its Applic., V. 463. Kluwer Academic, 1999, xvi + 648 pp, $295. [ISBN 0-7923-5504-0]

Analysis, P. *Partial Differential and Inte-*

gral Equations. Eds: Heinrich G.W. Begehr, Robert P. Gilbert, Guo-Chun Wen. Intern. Soc. for Analy., Applic. & Computat., V. 2. Kluwer Academic, 1999, x + 369 pp, $168. [ISBN 0-7923-5482-6] Proceedings of a 1997 congress. Topics: function theoretic and functional analytic methods for PDEs; applications of function theory of several complex variables to PDEs; integral equations and boundary value problems; PDEs.

Algebraic Geometry, P. *Mirror Symmetry and Algebraic Geometry.* David A. Cox, Sheldon Katz. Math. Surveys & Mono., V. 68. AMS, 1999, xxi + 469 pp, $69. [ISBN 0-8218-1059-6] Supersymmetric string theories exhibit interesting properties yet have shaky foundations, unverified by experiment. Exposits the algebraic geometry likely needed for a firm mathematical foundation. "Mirror pairs" of Calabi–Yau manifolds play a central role. RM

Algebraic Geometry, P. *The Curves Seminar at Queen's, Volume XII.* Ed: Anthony V. Geramita. Pure & Appl. Math., V. 114. Queen's Univ, 1998, 162 pp, (P). [ISBN 0-88911-832-9] Includes an expository paper on "Computational Invariant Theory" by Gregor Kemper as well as five contributed notes.

Differential Geometry, T?(17–18: 1), P. *Meromorphic Functions and Projective Curves.* Kichoon Yang. Math. & Its Applic., V. 464. Kluwer Academic, 1999, viii + 201 pp, $105. [ISBN 0-7923-5505-9] Concise, self-contained introduction and exposition of meromorphic functions of algebraic curves from a geometric viewpoint. Topics include Brill–Noether theory, projective differential geometry, minimal surfaces in Kahler manifolds. RM

Topology, T(17: 2), P, L. *Aspects of Topology, Second Edition.* Charles O. Christenson, William L. Voxman. BCS Associates, 1998, x + 493 pp, $48 (P); $75. [ISBN 0-914351-08-7; 0-914351-07-9] Topics include topological spaces, continua, homotopy theory, n-manifolds, and dimension theory. Changes from *First Edition* (TR, April 1978): rewrites on CW-complexes and covering spaces, various clarifications. JD

Operations Research, T(17–18: 1), P. *Geometric Methods and Optimization Problems.* V. Boltyanski, H. Martini, V. Soltan. Comb. Optim., V. 4. Kluwer Academic, 1999, viii + 429 pp, $204. [ISBN 0-7923-5454-0] Applies methods from convex geometry to solve optimization problems in control theory (modern variational geometry), location science, and computational geometry. Many intuitive diagrams motivate the ideas. RM

Optimization, T*(15–16: 1). *Practical Genetic Algorithms.* Randy L. Haupt, Sue Ellen Haupt. Wiley, 1998, xiv + 177 pp, $44.95. [ISBN 0-471-18873-5] Genetic algorithms use the model of evolutionary selection to solve a wide variety of optimization problems. Text starts with the basics of optimization and then shows how both discrete and continuous problems can be attacked using genetic algorithms. Many applications, some pseudocode, good references. No exercises. All in all, an interesting book. MPR

Optimization, P. *Minimax Theorems and Qualitative Properties of the Solutions of Hemivariational Inequalities.* D. Motreanu, P.D. Panagiotopoulos. Nonconvex Optim. & Its Applic., V. 29. Kluwer Academic, 1999, xviii + 309 pp, $156. [ISBN 0-7923-5456-7]

Optimization, P. *Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods.* Eds: Masao Fukushima, Liqun Qi. Appl. Optim., V. 22. Kluwer Academic, 1999, viii + 441 pp, $194. [ISBN 0-7923-5320-X] 22 refereed papers; most based on talks given at the 16th International Symposium on Mathematical Programming held at Lausanne, EPFL, Switzerland, in 1997.

Elementary Statistics, T(17: 1), C. *Data, Statistics, and Decision Models with Excel.* Donald L. Harnett, James F. Horrell. Wiley, 1998, xviii + 605 pp, $93.95, with disks. [ISBN 0-471-13398-1] Conceptual modeling and statistics text for MBA students. Emphasizes problem solving and decision making; calculus not required. Integrates use of Excel. Disks contain data and an Excel add-in. HS

Elementary Statistics, T(14: 2). *Introduction to Probability and Statistics for Scientists and Engineers.* Walter A. Rosenkrantz. Ser. in Prob. & Stat. McGraw-Hill, 1997, xiv + 592 pp. [ISBN 0-07-053988-X] Covers basic probability theory, sampling distributions, estimation, inference, linear and multiple regression, ANOVA, block designs, and control charts. MPR

Mathematical Statistics, T(15: 1). *Mathematical Statistics: An Introduction.* Wiebe R. Pestman. Walter de Gruyter, 1998, ix + 545 pp, $79 (P). [ISBN 3-11-015356-4] Theoretically rigorous introduction; includes a chapter on probability theory. Assumes knowledge of calculus and linear algebra. Companion volume (see following review) contains solutions to all 260 exercises. HS

Mathematical Statistics, S. *Mathematical Statistics: Problems and Detailed Solutions.*

Wiebe R. Pestman, Ivo B. Alberink. Walter de Gruyter, 1998, ix + 325 pp, $79 (P). [ISBN 3-11-015358-0] Detailed solutions to all 260 exercises in Pestman's *Mathematical Statistics: An Introduction* (see previous review). HS

Statistical Methods, T(18:2), P. *Statistical Learning Theory.* Vladimir N. Vapnik. Wiley, 1998, xxiv + 736 pp, $105. [ISBN 0-471-03003-1] Statistical learning theory explores ways of estimating functional dependency from a given collection of data for small samples without *a priori* knowledge about the problem to be solved. This comprehensive study of learning processes contains: general qualitative theory including necessary and sufficient conditions for consistency; general quantitative theory including bounds on the rate of convergence; principles for estimating functions from small data sets; methods of function estimation; applications to real-life problems. KB

Statistical Methods, T(16–17: 1), P. *Sampling of Populations: Methods and Applications, Third Edition.* Paul S. Levy, Stanley Lemeshow. Ser. in Prob. & Stat. Wiley, 1999, xxxi + 525 pp, $89.95. [ISBN 0-471-15575-6] Refines *Second Edition*'s treatment of telephone sampling and interviewing methodology. Material on survey data analysis now discusses use of appropriate software. HS

Statistical Methods, P. *The Analysis of Variance.* Henry Scheffé. Wiley Classics Library. Wiley, 1999, xvi + 477 pp, $49.95 (P). [ISBN 0-471-34505-9] Paperback printing of text originally published in 1959.

Statistics, C. *JMP Start Statistics: A Guide to Statistics and Data Analysis Using JMP and JMP IN Software.* John Sall, Ann Lehman. Duxbury Pr (Wadsworth), 1996, xxii + 521 pp, $61.95 (P), with disks. [ISBN 0-534-26565-0] Detailed introduction to JMP and JMP IN. Step-by-step instructions on how to perform many common statistical tests and procedures. Very thorough; many screen shots. MPR

Statistics, P. *Entropy Methods in Statistical Estimation.* M.H. Wegkamp. CWI Tract, V. 125. Centrum voor Wiskunde en Informatica, 1998, 120 pp, Dfl. 35 (P). [ISBN 90-6196-483-0]

Mathematical Computing, T(15–16), S. *Advanced Engineering Mathematics with Mathematica and MATLAB, Volume 1.* Reza Malek-Madani. Addison-Wesley, 1998, xiv + 558 pp, (P). [ISBN 0-201-59881-7] After brief introductions to Mathematica and MATLAB, covers ODEs, transform methods, linear algebra, systems of DEs, and numerical methods. Nice introduction to the use of these packages. Plenty of exercises. MPR

Computer Science, T(18), P. *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications.* Gunter Bolch, *et al.* Wiley, 1998, xvi + 726 pp, $89.95. [ISBN 0-471-19366-6] Presents the theory of computer performance analysis from the perspective of queueing theory and Markov chains. Applications include client-server systems, polling systems, operating systems, ATM networks. MPR

Applications (Biological Science), P. *Mathematical and Computational Biology: Computational Morphogenesis, Hierarchical Complexity, and Digital Evolution.* Ed: Chrystopher L. Nehaniv. Lect. on Math. in the Life Sci., V. 26. AMS, 1999, xi + 201 pp, $59 (P). [ISBN 0-8218-0941-5] Proceedings of a 1997 workshop at the University of Aizu, Japan.

Applications (Physics), T(17: 2), P. *Geometry, Particles, and Fields.* Bjørn Felsager. Grad. Texts in Contemp. Physics. Springer-Verlag, 1998, x + 672 pp, $69.95. [ISBN 0-387-98267-1] Geometric approach to modern particle field theory. First half is a relatively self-contained introduction to field theory, with an emphasis on gauge theory and nonlinear field theory. Second half is an introduction to differential geometry and its application to field theory. MPR

Applications (Statistical Mechanics), T*(16–17: 2), P*. *A Modern Course in Statistical Physics, Second Edition.* L.E. Reichl. Wiley, 1998, xix + 822 pp, $84.95. [ISBN 0-471-59520-9] Thorough, self-contained introduction to statistical mechanics using thermodynamics and probability theory as a foundation. Especially interesting for its extended coverage of non-equilibrium processes. Excellent examples and exercises. An outstanding choice for either a text or a reference. MPR

Applications, P. *Traffic Control and Transport Planning: A Fuzzy Sets and Neural Networks Approach.* Dušan Teodorović, Katarina Vukadinović. Intern. Ser. in Intelligent Tech. Kluwer Academic, 1998, xviii + 387 pp, $150. [ISBN 0-7923-8380-X]

Reviewers

KB: Karla Ballman, Macalester; DB: David Bressoud, Macalester; JD: Jill Dietz, St. Olaf; PF: Paul Froeschl, Macalester; PG: Philip Gloor, St. Olaf; BH: Bruce Hanson, St. Olaf; LCL: Loren C. Larson, St. Olaf; RM: Richard Molnar, Macalester; AO: Arnold Ostebee, St. Olaf; MPR: Matthew P. Richey, St. Olaf; KS: Karen Saxe, Macalester; HS: Heidi Shierholz, St. Olaf; LAS: Lynn Arthur Steen, St. Olaf; PZ: Paul Zorn, St. Olaf.

LESTER R. FORD AWARDS FOR 1998

The Lester R. Ford Awards, established in 1964, are made annually to authors of outstanding expository papers in the MONTHLY. The awards are named for Lester R. Ford, Sr., a distinguished mathematician, editor of the MONTHLY (1942–46), and President of the Mathematical Association of America (1947–1948).

Winners of the Lester R. Ford Awards for expository papers appearing in Volume 105 (1998) of the MONTHLY are:

Yoav Benyamini, Technion–Israel Institute of Technology, Applications of the Universal Surjectivity of the Cantor Set, pp. 832–839

A classical theorem due to Alexandroff and Hausdorff states that every compact metric space is the continuous image of the Cantor set. In this paper Yoav Benyamini presents striking applications of this result to diverse areas of mathematics. Each of these applications involves an existence theorem that Benyamini shows us how to prove using the universal surjectivity of the Cantor set. Some of these results are well known, such as the existence of space-filling curves and the isometric identification of every separable Banach space with a subspace of $C([0,1])$. Other results are more unusual, such as the existence of a compact convex subset of $\mathbf{R}^{n+2}$ whose faces include congruent copies of all compact convex subsets of the unit cube in $\mathbf{R}^n$. Other results are even more counter-intuitive, such as the existence of a continuous real-valued function f on $\mathbf{R}$ with the property that for every bounded sequence (a_n) of real numbers, there exists $t \in \mathbf{R}$ with $f(t + n) = a_n$ for all n. Benyamini ties all these results together in a pretty package with the common theme that the Cantor set and its universal surjectivity lurk behind many strange phenomena.

Jerry L. Kazdan, University of Pennsylvania, Solving Equations, an Elegant Legacy, pp. 1–21

The paper discusses various types of equations: polynomial equations in one and several variables, linear and nonlinear differential equations, diophantine equations, and congruences. The overriding idea is that familiar procedures for solving equations, often viewed as "tricks", can be seen as belonging to broad themes that, in turn, yield new insights on equations. Among the themes are: exploiting symmetry, finding a related problem, understanding the family of all solutions, finding obstructions when an equation has no solution, using variational methods, and reformulating a problem. An extensive discussion of symmetry is an important unifying thread. It bears on complex conjugation, linear differential equations, Markov chains, Lie's "Galois" theory of differential equations, and Pell's equation. Kazdan's article is an instructive and wide-ranging tour of the mathematician's workshop in important classes of equations.

Bernd Sturmfels, University of California, Berkeley, Polynomial Equations and Convex Polytopes, pp. 907–922

How many complex zeros can d polynomials in d variables have? In the case of the bivariate system

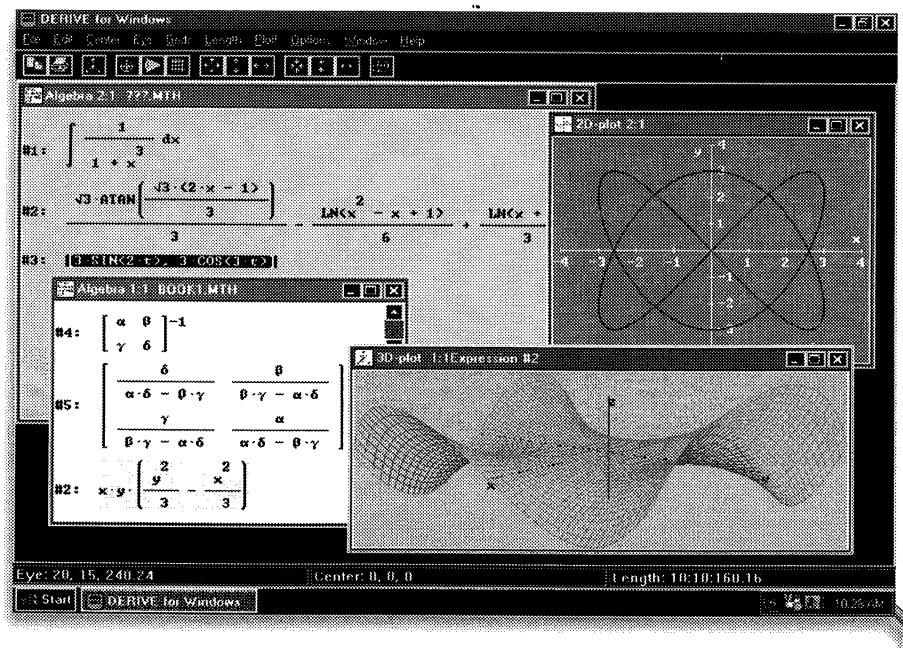
$$a_1 + a_2 x + a_3 xy + a_4 y = b_1 + b_2 x^2 y + b_3 xy^2 = 0$$

with nonzero real coefficients, Bezout's theorem gives an upper bound of six solutions. But it has exactly four. To achieve this better estimate, we use an idea of Newton. To a bivariate polynomial $\Sigma x^u y^v$, associate its *Newton polygon*, the convex hull of the vertices (u, v). The mixed area $\mathcal{M}(P, Q)$ of two planar polygons P, Q is defined by

$$\mathcal{M}(P, Q) = \text{area}(P + Q) - \text{area}(P) - \text{area}(Q).$$

In 1975, David Bernstein proved a general theorem that for two equations in two unknowns shows that the number of solutions of a system of two bivariate polynomial equations is equal to the mixed area of the two corresponding Newton polygons. Sturmfels outlines an algorithmic proof devised by B. Huber and himself in 1995 that leads to a numerical approximation for the solution. The author deftly avoids getting overwhelmed by algebraic and geometric detail by using examples and organizing his account of the proof of Bernstein's theorem around three key steps. The case of real zeros has been seriously investigated for only the past twenty years and little is known. Sturmfels brings us into the cut and thrust of current research with its distance between conjecture and reality and its open questions, leaving much to do that is of interest to combinatorialists, algebraic geometers, and applied mathematicians.

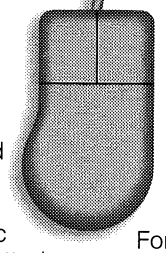
NEW! ***Site Licenses and Student Pricing. See www.derive.com***



DERIVE® for Windows

DERIVE is the trusted mathematical assistant relied upon by students, educators, engineers, and scientists around the world. It does for algebra, equations, trigonometry, vectors, matrices, and calculus what the scientific calculator does for numbers — it eliminates the drudgery of performing long and tedious mathematical calculations. You can easily solve both symbolic and numeric problems and see the results plotted as 2D or 3D graphs.

For everyday mathematical work *DERIVE* is a tireless, powerful, and knowledgeable assistant. For teaching or learning mathematics, *DERIVE* gives you the freedom to explore different mathematical approaches better and more quickly than by using traditional methods.



System Requirements: Windows 95, 3.1x or NT running on a computer with 8 megabytes of memory.

Suggested Retail Price: $250. Educational pricing available.

For product information and list of dealers, fax, email, write, or call Soft Warehouse, Inc. or visit our website at http://www.derive.com.

The Easiest just got Easier.

Soft Warehouse INC
HONOLULU • HAWAII

Soft Warehouse, Inc. • 3660 Waialae Avenue
Suite 304 • Honolulu, Hawaii, USA 96816-3259
Telephone: (808) 734-5801 after 10:00 a.m. PST
Fax: (808) 735-1105 • Email: swh@aloha.com.